

# Decentralized Projection-free Optimization for Convex and Non-convex Problems

Hoi-To Wai, Jean Lafond, Anna Scaglione, Eric Moulines <sup>\*†</sup>

December 6, 2016

## Abstract

Decentralized optimization algorithms have received much attention as fueled by the recent advances in network information processing and the tremendous amount of data that is generated by human activities. However, conventional decentralized algorithms based on projected gradient descent are incapable of handling high dimensional constrained problems, as the projection step becomes computationally prohibitive to compute. To address this problem, we adopt a projection-free optimization approach, a.k.a. the Frank-Wolfe (FW) or conditional gradient algorithm. We first develop a decentralized FW (DeFW) algorithm from the classical FW algorithm. The convergence of the proposed algorithm is studied by viewing the decentralized algorithm as an *inexact* FW algorithm. Using a diminishing step size rule and letting  $t$  be the iteration number, we show that the DeFW algorithm's convergence rate is  $\mathcal{O}(1/t)$  for convex objectives; is  $\mathcal{O}(1/t^2)$  for strongly convex objectives with the optimal solution in the interior of the constraint set; and is  $\mathcal{O}(1/\sqrt{t})$  towards a stationary point for smooth but non-convex objectives. We then show that a gossip-based implementation meets the above guarantees with two communication rounds per iteration. Furthermore, we demonstrate the advantages of the proposed DeFW algorithm on two applications including low-complexity robust matrix completion and communication efficient sparse learning. Numerical results on synthetic and realistic data are presented to support our findings.

## 1 Introduction

Recently, algorithms for tackling high-dimensional optimizations have been sought for handling the large volume of data [3] generated by human activities. Since these data are dispersed over clouds of networked computers, it is important to consider *decentralized* algorithms that can allow the cloud/agents/nodes to co-operate and share computational resources across the network [4].

---

<sup>\*</sup>The work of H.-T. Wai is supported by NSF CCF-1011811 and J. Lafond by Direction Générale de l'Armement and the labex LMH (ANR-11-LABX-0056-LMH). Preliminary versions of this work are presented at ICASSP 2016, Shanghai, China [1] and GlobalSIP 2016, Washington DC, USA [2].

<sup>†</sup>H.-T. Wai and A. Scaglione are with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA. E-mails: {htwai, Anna.Scaglione}@asu.edu. J. Lafond was with Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France. Email: lafond.jean@gmail.com. E. Moulines is with CMAP, Ecole Polytechnique, Palaiseau, France. Email: eric.moulines@polytechnique.edu.

This paper considers decentralized algorithm for tackling a constrained optimization problem whose objective function can be expressed as an average of  $N$  functions, i.e.,

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} F(\boldsymbol{\theta}) \text{ s.t. } \boldsymbol{\theta} \in \mathcal{C}, \text{ where } F(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^N f_i(\boldsymbol{\theta}), \quad (1)$$

where  $f_i(\boldsymbol{\theta})$  is a continuously differentiable (possibly non-convex) function held privately by the  $i$ th agent and  $\mathcal{C}$  is a closed and bounded convex set in  $\mathbb{R}^d$ . Typically  $\mathcal{C}$  corresponds to some constraints imposed on the parameter  $\boldsymbol{\theta}$  such as sparsity or low rank of an array of values. In fact, Problem (1) covers a number of applications in signal processing and machine learning, such as matrix completion [5] and distributed LASSO [6, 7]. As a standard assumption in decentralized optimization, the private functions  $f_i(\boldsymbol{\theta})$  are not sharable among the agents due to data privacy and dimensionality concerns. The agents communicate on a network that is described by a graph  $G = (V, E)$ , where  $V = \{1, \dots, N\}$  is the set of agents and  $E \subseteq V \times V$  describes the connectivity.

As  $G$  is not fully connected, it is useful to apply decentralized algorithms capable of performing *in-network* computations. Various authors proposed decentralized optimization algorithms to solve (1) that are built on the gossip-based average consensus protocol [8, 9]. Examples of prior works include [10, 11] which studied the decentralized counterparts of projected gradient descent (PGD) methods; [12, 13] which considered the decentralized primal-dual/ADMM algorithms; [14, 15] which considered the successive convex approximation methods. The convergence properties and the performance of these algorithms were investigated extensively, especially when the objective is convex, see [4, 10–12, 16–19]; for non-convex objectives, some recent results have been reported in [13, 15, 20–22]. A common trait found in the existing methods on the subject is that each iteration of these algorithms require at least one *projection* operation onto the constraint set  $\mathcal{C}$ . When the size of the problem is moderate and  $\mathcal{C}$  is structured, this projection step can be computed efficiently. When the problem involves a high dimensional parameter, i.e.,  $d \gg 0$ , the projection step may be computationally prohibitive, rendering most existing methods impractical.

This paper focuses on developing a decentralized *projection-free* optimization algorithm. Specifically, we extend the Frank-Wolfe (FW) algorithm [23] to operate in a decentralized setting. The FW algorithm has been recently popularized due to its efficacy in handling high-dimensional constrained problems. Examples of its applications include matrix completion [24], image and video colocation [25], electric vehicle charging optimization [26] and traffic assignment [27]; see the overview article [28]. From the algorithmic perspective, the FW algorithm replaces the costly projection step in PGD-based algorithms with a constrained linear optimization, which often admits an efficient solution, e.g., when the constraint set is a trace-norm ball or a polytope.

Our contributions are as follows. We first describe abstractly the proposed algorithm as a variation of FW algorithm operating on *inexact* iterates and gradients, with the latter operations computed by in-network operations. We then analyze its convergence — for convex objectives, the sub-optimality (of the objective values) of the iterates produced by the proposed algorithm is shown to converge as  $\mathcal{O}(1/t)$  with  $t$  being the iteration number, and is  $\mathcal{O}(1/t^2)$  for strongly convex objectives if the optimal solution is in the interior of  $\mathcal{C}$ ; for non-convex objectives, we demonstrate that the proposed algorithm has limit points that are stationary points of (1), and they can be found at the rate of  $\mathcal{O}(\sqrt{1/t})$ . We show that a gossip-based implementation of the proposed algorithm with fixed number of communication rounds achieves all the above guarantees. To our knowledge, this is the first application of FW algorithm in a decentralized setting and the convergence rate in the non-convex setting is comparable to that of a *centralized* PG method [29]. Lastly, we

present examples where the proposed algorithm can be applied, including a communication-efficient decentralized LASSO solver and decentralized matrix completion.

The rest of this paper is organized as follows. Section 2 develops the DeFW algorithm. We summarize the main theoretical results of convergence for convex and non-convex objective functions. Several implementation related issues will also be discussed. A gossip-based implementation of the DeFW algorithm will then be presented in Section 3. Applications of the DeFW algorithm are discussed in Section 4. Finally, in Section 5, numerical results are shown to support our theoretical findings.

## 1.1 Notations & Mathematical Preliminaries

For any  $d \in \mathbb{N}$ , we define  $[d]$  as the set  $\{1, \dots, d\}$ . We use boldface lower-case letters to denote the vectors and boldfaced upper-case letters to denote matrices. For a vector  $\mathbf{x}$  (or a matrix  $\mathbf{X}$ ), the notation  $[\mathbf{x}]_i$  (or  $[\mathbf{X}]_{i,j}$ ) denotes its  $i$ th element (or  $(i, j)$ th element). The vectorization of a matrix  $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$  is denoted by  $\text{vec}(\mathbf{X}) = [\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_{m_2}] \in \mathbb{R}^{m_1 m_2}$  such that  $\mathbf{x}_i$  is the  $i$ th column of  $\mathbf{X}$ . The vector  $\mathbf{e}_i \in \mathbb{R}^d$  is the  $i$ th unit vector such that  $[\mathbf{e}_i]_j = 0$  for all  $j \neq i$  and  $[\mathbf{e}_i]_i = 1$ . For some positive finite constants  $C_1, C_2, C_3, C_4$  with  $C_3 \leq C_4$  and non-negative functions of  $t$ ,  $f(t), g(t)$ , the notations  $f(t) = \mathcal{O}(g(t))$ ,  $f(t) = \Omega(g(t))$  and  $f(t) = \Theta(g(t))$  indicates that  $f(t) \leq C_1 g(t)$ ,  $f(t) \geq C_2 g(t)$  and  $C_3 g(t) \leq f(t) \leq C_4 g(t)$  for sufficiently large  $t$ , respectively.

Let  $\mathbf{E}$  be a Euclidean space embedded in  $\mathbb{R}^d$  and the associated Euclidean norm is denoted by  $\|\cdot\|_2$ . The binary operator  $\langle \cdot, \cdot \rangle$  denotes the inner product on  $\mathbf{E}$ . In addition,  $\mathbf{E}$  is equipped with a norm  $\|\cdot\|$  and the corresponding dual norm  $\|\cdot\|_\star$ . Let  $G, L, \mu$  be some non-negative constants. Consider a function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ , the function  $f$  is said to be  $G$ -Lipschitz if for all  $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{E}$

$$|f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}')| \leq G \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\star; \quad (2)$$

the function  $f$  is  $L$ -smooth if for all  $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{E}$

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \leq \langle \nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle + \frac{L}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \quad (3)$$

note that the above is equivalent to  $\|\nabla f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta})\|_2 \leq L \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2$ ; the function  $f$  is  $\mu$ -strongly convex if for all  $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{E}$ ,

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \leq \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle - \frac{\mu}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \quad (4)$$

notice that if  $\mu = 0$ , the above definition reduces to that of stating that  $f$  is convex.

Consider Problem (1), its constraint set  $\mathcal{C} \subseteq \mathbf{E}$  is convex and bounded with the diameter defined as:

$$\rho := \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\star, \quad \bar{\rho} := \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \quad (5)$$

note that  $\rho$  is defined with respect to (w.r.t.) the dual norm  $\|\cdot\|_\star$  while  $\bar{\rho}$  is defined w.r.t. the Euclidean norm. When the objective function  $F$  is  $\mu$ -strongly convex with  $\mu > 0$ , the optimal solution to (1) is unique and denoted by  $\boldsymbol{\theta}^\star$ , we also define

$$\delta := \inf_{\mathbf{s} \in \partial \mathcal{C}} \|\mathbf{s} - \boldsymbol{\theta}^\star\|_2, \quad (6)$$

where  $\partial \mathcal{C}$  is the boundary set of  $\mathcal{C}$ . If  $\delta > 0$ , the solution  $\boldsymbol{\theta}^\star$  is in the interior of  $\mathcal{C}$ .

---

**Algorithm 1** Decentralized Frank-Wolfe (DeFW).

---

- 1: **Input:** Initial point  $\theta_1^i$  for  $i = 1, \dots, N$ .
- 2: **for**  $t = 1, 2, \dots$  **do**
- 3:   *Consensus:* obtain the average parameter:

$$\bar{\theta}_t^i \leftarrow \text{NetAvg}_t^i(\{\theta_t^j\}_{j=1}^N), \quad \forall i \in [N].$$

- 4:   *Aggregating:* obtain the average gradient:

$$\overline{\nabla_t^i F} \leftarrow \text{NetAvg}_t^i(\{\nabla f_j(\bar{\theta}_t^j)\}_{j=1}^N), \quad \forall i \in [N].$$

- 5:   *Frank-Wolfe Step:* update

$$\theta_{t+1}^i \leftarrow (1 - \gamma_t)\bar{\theta}_t^i + \gamma_t \mathbf{a}_t^i \quad \text{where} \quad \mathbf{a}_t^i = \arg \min_{\theta \in \mathcal{C}} \langle \overline{\nabla_t^i F}, \theta \rangle,$$

for all agent  $i \in [N]$  and  $\gamma_t \in (0, 1]$  is a step size.

- 6: **end for**

- 7: **Return:**  $\bar{\theta}_{t+1}^i, \forall i \in [N]$ .
- 

## 2 Decentralized Frank-Wolfe (DeFW)

We develop the decentralized Frank-Wolfe (DeFW) algorithm from the classical Frank-Wolfe (FW) algorithm [23]. A main advantage of the FW algorithm is that it is *projection-free*. In particular, let  $t \in \mathbb{N}, t \geq 1$  be the iteration number and the initial point  $\theta_0 \in \mathcal{C}$  be feasible, the *centralized* FW algorithm for (1) proceeds by:

$$\theta_t = \theta_{t-1} + \gamma_{t-1}(\mathbf{a}_{t-1} - \theta_{t-1}), \quad (7a)$$

$$\mathbf{a}_{t-1} = \arg \min_{\mathbf{a} \in \mathcal{C}} \langle \nabla F(\theta_{t-1}), \mathbf{a} \rangle. \quad (7b)$$

The scalar  $\gamma_t \in (0, 1]$  is a step size to be determined. Observe that  $\theta_t$  is a convex combination of  $\theta_{t-1}$  and  $\mathbf{a}_{t-1}$ , therefore  $\theta_t \in \mathcal{C}$  since  $\mathcal{C}$  is a convex set. Here, the linear optimization (LO) (7b) can often be solved efficiently depending on the structure of  $\mathcal{C}$ , as we remark at the end of this section. Moreover, when we choose the step size as  $\gamma_t = 2/(t+1)$ , the FW algorithm is known to converge at a rate of  $\mathcal{O}(1/t)$  if  $F$  is  $L$ -smooth and convex [28].

Our plan is to extend the FW algorithm to a decentralized setting via mimicking (7) with in-network operations. We first offer a high-level description of the proposed DeFW algorithm — let  $\theta_t^i$  denotes the parameter kept by agent  $i$  at iteration  $t$ . Define the average parameter:

$$\bar{\theta}_t := N^{-1} \sum_{i=1}^N \theta_t^i. \quad (8)$$

Also, we denote  $\bar{\theta}_t^i$  as an approximate of the network average parameter. To mimic the FW algorithm (7), we require the local approximate average  $\bar{\theta}_t^i$  to trace the network average  $\bar{\theta}_t$  with an increasing accuracy. We assume that:

**H1** Let  $\{\Delta p_t\}_{t \geq 1}$  be a non-negative, decreasing sequence, we have

$$\max_{i \in [N]} \|\bar{\theta}_t^i - \bar{\theta}_t\|_2 \leq \Delta p_t, \quad \forall t \geq 1. \quad (9)$$

To compute the LO, each agent has to have access to the global gradient  $\nabla F(\bar{\theta}_t)$ . However from the local average parameter  $\bar{\theta}_t^i$ , only the local gradient  $\nabla f_i(\bar{\theta}_t^i)$  is available. To this end, we let  $\overline{\nabla}_t^i F$  be the approximate of the global gradient kept by agent  $i$ , that traces the network average gradient,

$$\overline{\nabla}_t F := N^{-1} \sum_{j=1}^N \nabla f_j(\bar{\theta}_t^j). \quad (10)$$

We assume that:

**H2** Let  $\{\Delta d_t\}_{t \geq 1}$  be a non-negative, decreasing sequence, we have

$$\max_{i \in [N]} \|\overline{\nabla}_t^i F - \overline{\nabla}_t F\|_2 \leq \Delta d_t, \quad \forall t \geq 1. \quad (11)$$

Naturally, the  $i$ th agent can locally compute the update direction  $\mathbf{a}_t^i := \arg \min_{\theta \in \mathcal{C}} \langle \overline{\nabla}_t^i F, \theta \rangle$  using the local approximation  $\overline{\nabla}_t^i F$ . We can summarize the proposed DeFW algorithm in Algorithm 1. Notice that the subroutine  $\text{NetAvg}_t^i(\cdot)$  indicates that the vectors  $\bar{\theta}_t^i$  and  $\overline{\nabla}_t^i F$  are both to be computed using in-network operations, e.g., the gossip-based consensus protocol described in Section 3.

Under H1 and H2, for each agent  $i \in [N]$ , line 5 in the DeFW algorithm can be regarded as performing the FW updates (7) on  $\bar{\theta}_t$  in an inexact manner. Below we characterize the convergence of the DeFW algorithm. For convex objective functions, we have:

**Theorem 1** Set the step size as  $\gamma_t = 2/(t+1)$ . Suppose that each of  $f_i$  is convex and  $L$ -smooth. Under H1, H2 with  $\Delta p_t = C_p/t$ ,  $\Delta d_t = C_g/t$ , we have

$$F(\bar{\theta}_t) - F(\theta^*) \leq \frac{8\bar{\rho}(C_g + LC_p) + 2L\bar{\rho}^2}{t+1}, \quad (12)$$

for all  $t \geq 1$ , where  $\theta^*$  is an optimal solution to (1). Furthermore, if  $F$  is  $\mu$ -strongly convex and the optimal solution  $\theta^*$  lies in the interior of  $\mathcal{C}$ , i.e.,  $\delta > 0$  (cf. (6)), we have

$$F(\bar{\theta}_t) - F(\theta^*) \leq \frac{(4\bar{\rho}(C_g + LC_p) + L\bar{\rho}^2)^2}{2\delta^2\mu} \cdot \frac{9}{(t+1)^2}, \quad (13)$$

for all  $t \geq 1$ .

The proof can be found in Appendix A. Since  $F(\theta)$  is convex in the above, the conditions (12), (13) imply that the iterate sequence  $\{\bar{\theta}_t\}_{t \geq 1}$  converges to an optimal solution of (1).

For non-convex objective, we study the convergence of the FW/duality gap, defined as  $g_t := \max_{\theta \in \mathcal{C}} \langle \nabla F(\bar{\theta}_t), \bar{\theta}_t - \theta \rangle$ . Also, define the set of stationary point to (1) as:

$$\mathcal{C}^* = \{\underline{\theta} \in \mathcal{C} : \max_{\theta \in \mathcal{C}} \langle \nabla F(\underline{\theta}), \underline{\theta} - \theta \rangle = 0\}, \quad (14)$$

and we assume that

**H3** The function  $F(\theta)$  takes a finite number of values in  $\mathcal{C}^*$ , i.e., the image set  $F(\mathcal{C}^*) = \{F(\theta) : \theta \in \mathcal{C}^*\}$  is finite.

**Theorem 2** Set the step size as  $\gamma_t = 1/t^\alpha$  for some  $\alpha \in (0, 1]$ . Suppose that each of  $f_i$  is  $L$ -smooth and  $G$ -Lipschitz (possibly non-convex). Under H1, H2 with  $\Delta p_t = C_p/t^\alpha$ ,  $\Delta d_t = C_g/t^\alpha$  and  $C_p, C_g$  are finite, then —

1. The following hold for all  $T \geq 6$  that are even, if  $\alpha \in [0.5, 1)$ , we have

$$\min_{t \in [T/2+1, T]} g_t \leq \frac{1}{T^{1-\alpha}} \cdot \frac{1-\alpha}{(1-(2/3)^{1-\alpha})} \cdot \left( G\rho + (L\bar{\rho}^2/2 + 2\bar{\rho}(C_g + LC_p)) \log 2 \right); \quad (15)$$

if  $\alpha \in (0, 0.5)$ , we have

$$\min_{t \in [T/2+1, T]} g_t \leq \frac{1}{T^\alpha} \cdot \frac{1-\alpha}{(1-(2/3)^{1-\alpha})} \cdot \left( G\rho + \frac{(L\bar{\rho}^2/2 + 2\bar{\rho}(C_g + LC_p))(1-(1/2)^{1-2\alpha})}{1-2\alpha} \right). \quad (16)$$

2. Under H3, if  $\alpha \in (0.5, 1]$ , then the sequence of objective values  $\{F(\bar{\theta}_t)\}_{t \geq 1}$  converges, the sequence  $\{\bar{\theta}_t\}_{t \geq 1}$  has limit points and each of the limit point  $\underline{\theta}$  is in  $\mathcal{C}^*$ .

The proof can be found in Appendix B. From the definition, when the FW gap is zero,  $g_t = 0$ , then the iterate  $\bar{\theta}_t$  is a stationary point to (1). Therefore, we can regard  $g_t$  as a measure of the stationarity of the iterate  $\bar{\theta}_t$ . From (15), setting  $\alpha = 0.5$  gives the convergence rate of  $\min_{t \in [T/2+1, T]} g_t = \mathcal{O}(1/\sqrt{T})$ .

The theorems above require the consensus error,  $\max_{i \in [N]} \|\bar{\theta}_t^i - \bar{\theta}_t\|_2$ , to decay to zero as  $t \rightarrow \infty$ , as such, the local iterates  $\bar{\theta}_t^i$  converge to an optimal/stationary solution of (1). Before concluding this section, we remark that the DeFW algorithm requires each agent to solve the LO (7b) independently at each iteration. As we have mentioned, this can be done efficiently for several interesting cases of the constraint set  $\mathcal{C}$ . For example:

- When  $\mathcal{C}$  is the  $\ell_1$  ball, i.e.,  $\mathcal{C} = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq R\}$ ,

$$\mathbf{a}^t = -R \cdot \mathbf{e}_i, \text{ where } i = \arg \max_{j \in [d]} |[\nabla F(\theta_t)]_j|. \quad (17)$$

The solution above to (7b) corresponds to finding the coordinate of the gradient vector with the maximum magnitude. Importantly, this solution is only 1-sparse. Consequently, the  $t$ th iterate  $\theta^t$  is at most  $t$ -sparse. The worst-case complexity of computing  $\mathbf{a}^t$  is  $\mathcal{O}(d)$ ; in comparison, the worst-case complexity for the projection into an  $\ell_1$  ball is  $\mathcal{O}(d \log d)^1$ .

- When  $\mathcal{C}$  is the trace norm ball, i.e.,  $\mathcal{C} = \{\theta \in \mathbb{R}^{m_1 \times m_2} : \|\theta\|_{\sigma,1} \leq R\}$ , let  $\mathbf{u}_1, \mathbf{v}_1$  be the top left/right singular vector of  $\nabla F(\theta_t)$ , we have

$$\mathbf{a}^t = -R \cdot \mathbf{u}_1 \mathbf{v}_1^\top, \quad (18)$$

i.e., it can be computed as the principal component of  $\nabla F(\theta_t)$ . Importantly, at a target solution accuracy of  $\delta$ , the top singular vectors can be computed with a complexity of  $\mathcal{O}(\max\{m_1, m_2\} \log(1/\delta))$  using the power or Lanczos method [31] if  $\|\text{vec}(\nabla F(\theta_t))\|_0 = \mathcal{O}(\max\{m_1, m_2\})$ , while the projection counterpart requires a complexity of  $\mathcal{O}(\max\{m_1 m_2^2, m_2 m_1^2\} \log(1/\delta))$  for computing the full SVD of a  $m_1 \times m_2$  matrix.

The above examples are relevant to the two applications described in Section 4. More recently, efficient implementations are found when  $\mathcal{C}$  admits additional structure such as being the convex hull of all perfect matchings of a bipartite graph; see [28] for an overview.

<sup>1</sup>There exists a randomized, accelerated algorithm for projection in [30] with an *expected* complexity of  $\mathcal{O}(d)$ .

### 3 Gossip-based DeFW algorithm

To guarantee convergence for the DeFW algorithm in both Theorem 1 & 2, the averages' approximation errors  $\Delta p_t, \Delta d_t$  converge to zero at a polynomial rate of  $t$ . In this section, we enforce this by employing the gossip-based average consensus (GAC) protocol [8, 9] with a *fixed* number of update/communication rounds for the  $\text{NetAvg}_t^i(\cdot)$  subroutine.

Specifically, the following discussions are based on the static GAC; note that it is possible to extend the protocol to a randomized setting for time-varying networks (e.g., with random link failures), see [32]. To fix idea, assume that the graph  $G$  is undirected and we define a non-negative *weight matrix*  $\mathbf{W} \in \mathbb{R}_+^{N \times N}$  that describes the local communication between the  $N$  agents. The matrix satisfies that  $W_{ij} := [\mathbf{W}]_{ij} > 0$  if and only if  $(i, j) \in E$ . It is also symmetric and doubly stochastic, i.e.,  $\mathbf{W}\mathbf{1} = \mathbf{W}^\top \mathbf{1} = \mathbf{1}$ . We assume that  $|\lambda_2(\mathbf{W})| < 1$ , i.e., the second largest eigenvalue of  $\mathbf{W}$  is strictly less than one, note that the existence of such matrix  $\mathbf{W}$  is guaranteed if  $G$  is connected. For each round of the GAC update, the agents take a weighted average of the values from its neighbors according to  $\mathbf{W}$ . We first state the following fact regarding the non-negative weight matrix  $\mathbf{W}$ .

**Fact 1** *Let  $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d$  be a set of  $N$  vectors and  $\mathbf{x}_{avg} := N^{-1} \sum_{i=1}^N \mathbf{x}_i$  be their average. Suppose  $\mathbf{W}$  is a doubly stochastic, non-negative matrix. The output after performing one round of GAC update:*

$$\bar{\mathbf{x}}_i = \sum_{j=1}^N W_{ij} \cdot \mathbf{x}_j \quad (19)$$

*must satisfy*

$$\sqrt{\sum_{i=1}^N \|\bar{\mathbf{x}}_i - \mathbf{x}_{avg}\|_2^2} \leq |\lambda_2(\mathbf{W})| \cdot \sqrt{\sum_{i=1}^N \|\mathbf{x}_i - \mathbf{x}_{avg}\|_2^2}, \quad (20)$$

*where  $\lambda_2(\mathbf{W})$  is the second largest eigenvalue of  $\mathbf{W}$ .*

The fact above can be verified from linear algebra. As  $\lambda_2(\mathbf{W}) < 1$ , the above implies that each GAC update (19) moves the vectors closer to the network average. Repeatedly applying (20) shows the well known fact that GAC computes the network average at a geometric rate.

Let us consider the in-network computation of  $\bar{\boldsymbol{\theta}}_t^i$  in line 3 of the DeFW algorithm. Here, the  $\text{NetAvg}_t^i(\cdot)$  subroutine in the *consensus step* is implemented by:

$$\bar{\boldsymbol{\theta}}_t^i = \sum_{j=1}^N W_{ij} \cdot \boldsymbol{\theta}_t^j, \quad (21)$$

i.e., we perform one round of the GAC update. Since  $W_{ij} = 0$  if  $(i, j) \notin E$ , the above operation is implemented using message exchanges among the neighbors of agent  $i$ .

Now, for some  $\alpha \in (0, 1]$ , we define  $t_0(\alpha)$  as the smallest integer such that

$$\lambda_2(\mathbf{W}) \leq \left( \frac{t_0(\alpha)}{t_0(\alpha) + 1} \right)^\alpha \cdot \frac{1}{1 + (t_0(\alpha))^{-\alpha}}. \quad (22)$$

Notice that  $t_0(\alpha)$  is upper bounded by:

$$t_0(\alpha) \leq \lceil (|\lambda_2(\mathbf{W})|^{-1/(1+\alpha)} - 1)^{-1} \rceil. \quad (23)$$

The following lemma can be easily proven:

**Lemma 1** *Set the step size  $\gamma_t$  as  $1/t^\alpha$  in the DeFW algorithm for some  $\alpha \in (0, 1]$ , then  $\bar{\theta}_t^i$  in (21) satisfies H1 with*

$$\Delta p_t = C_p/t^\alpha, \quad \forall t \geq 1, \quad \text{where } C_p := (t_0(\alpha))^\alpha \cdot \sqrt{N} \bar{\rho}. \quad (24)$$

The proof is postponed to Appendix C, which relies on using the fact that  $\theta_t^i$  is a linear combination of  $\bar{\theta}_{t-1}^i$  and  $\mathbf{a}_{t-1}^i$ , i.e., iterates from the previous iteration. In particular,  $\bar{\theta}_{t-1}^i$  is already  $\mathcal{O}(1/(t-1)^\alpha)$ -close to the network average from the last iteration and the update direction  $\mathbf{a}_{t-1}^i$  has to be weighted by the decaying step size  $\gamma_{t-1}$ .

In comparison to what we were able to establish above, the in-network computation of  $\overline{\nabla_t^i F}$  in line 4 of the DeFW algorithm is less straightforward. Unlike the computation of  $\bar{\theta}_t$ , computing  $N^{-1} \sum_{i=1}^N \nabla f_i(\bar{\theta}_t^i)$  to an accuracy of  $\mathcal{O}(1/t^\alpha)$  by communicating the local gradient  $\nabla f_i(\bar{\theta}_t^i)$  requires  $\Omega(\log t)$  rounds of updates when the GAC protocol is employed. One of the main technical issues is that the local gradient  $\nabla f_i(\bar{\theta}_t^i)$  computed by the  $i$ th agent is in general different from the local gradient computed at the other agent, even when  $\bar{\theta}_t^i$  is close to  $\bar{\theta}_t^j$  for  $j \neq i$ .

We adopt an approach that is inspired by the fast stochastic average gradient (SAGA) method [33] which re-uses the gradient approximate  $\overline{\nabla_{t-1}^i F}$  from the last iteration. Specifically, define the following surrogate of local gradient at iteration  $t$ :

$$\nabla_t^i F := \overline{\nabla_{t-1}^i F} + \nabla f_i(\bar{\theta}_t^i) - \nabla f_i(\bar{\theta}_{t-1}^i), \quad (25)$$

for all  $i \in [N]$ . When  $t = 1$ , we set  $\nabla_1^i F = \nabla f_i(\bar{\theta}_1^i)$ . Similar to (21), the  $\text{NetAvg}_t^i(\cdot)$  subroutine in the *aggregating* step is implemented by:

$$\overline{\nabla_t^i F} = \sum_{j=1}^N W_{ij} \cdot \nabla_t^j F, \quad (26)$$

i.e., using just one round of the GAC update on  $\nabla_t^i F$ . Below we show that the average gradient is preserved by  $\nabla_t^i F$  and  $\overline{\nabla_t^i F}$  has an approximation error similar to Lemma 1:

**Lemma 2** *Set the step size  $\gamma_t$  as  $1/t^\alpha$  in the DeFW algorithm for some  $\alpha \in (0, 1]$ . Suppose that each of  $f_i$  is  $L$ -smooth,  $\bar{\theta}_t^i$  is updated according to (21), then  $\overline{\nabla_t^i F}$  in (26) satisfies*

$$N^{-1} \sum_{i=1}^N \nabla_t^i F = N^{-1} \sum_{i=1}^N \nabla f_i(\bar{\theta}_t^i), \quad \forall t \geq 1, \quad (27)$$

and H2 with

$$\Delta d_t = C_g/t^\alpha, \quad \forall t \geq 1, \quad (28)$$

where

$$C_g := (t_0(\alpha))^\alpha \cdot 2\sqrt{N}(2C_p + \bar{\rho})L. \quad (29)$$

The proof can be found in Appendix D. Similar intuition as in Lemma 1 was used in the proof. In particular, we observe that  $\overline{\nabla_{t-1}^i F}$  is  $\mathcal{O}(1/(t-1)^\alpha)$ -close to the network average  $\overline{\nabla_{t-1} F}$  from the previous iteration and  $\nabla f_i(\bar{\theta}_t^i) - \nabla f_i(\bar{\theta}_{t-1}^i)$  scales as  $\|\bar{\theta}_t^i - \bar{\theta}_{t-1}^i\|_2 \leq \Delta p_{t-1} = \mathcal{O}(1/(t-1)^\alpha)$ .

**Remark 1** *It is possible for the agents to repeat the updates in (21), (26) for multiple rounds. Mathematically, this is equivalent to replacing  $W_{ij}$  in the above mentioned equations by  $[\mathbf{W}^\ell]_{ij}$ . As  $|\lambda_2(\mathbf{W}^\ell)| = |\lambda_2(\mathbf{W})|^\ell$ , the constants  $t_0(\alpha), C_p, C_g$  can be greatly reduced.*

Finally, observe that the conditions on  $\Delta p_t, \Delta d_t$  required by Theorem 1 & 2 can be satisfied by the  $\text{NetAvg}_t^i(\cdot)$  subroutine implemented with the GAC protocol in (21) & (26). This results in the following corollary.



**Corollary 1** *The convergence guarantees in Theorem 1 & 2 hold when the  $\text{NetAvg}_t^i(\cdot)$  subroutine in line 3, line 4 of the DeFW algorithm are implemented by (21), (26) respectively.*

In other words, the gossip-based DeFW algorithm converges for both convex and non-convex problems, while using a constant number of communication rounds per iteration. Lastly, the DeFW algorithm is not limited to the gossip-based implementation. In fact, any average consensus protocols which produce in-network averages satisfying H1, H2 with the desirable rates on  $\Delta p_t, \Delta d_t$  can be applied. For example, when the graph  $G$  of the communication network is directed, one may apply the average consensus algorithm in [34].

**Remark 2** *As we recall from Theorem 2, for non-convex objectives, the best theoretical rate of convergence can be achieved when if we set  $\alpha = 0.5$  as the learning rate. However, from Lemma 1 & 2, we notice that the approximation error also decays the slowest when  $\alpha = 0.5$ . From our numerical experience, we find that the approximation error  $\Delta p_t, \Delta d_t$  indeed play an important role in the practical performance of the DeFW algorithm. Therefore, we shall set  $\alpha$  to be higher than 0.5 for better performance.*

## 4 Applications of DeFW Algorithm

In this section, we study two applications of the DeFW algorithm in signal processing and machine learning problems.

### 4.1 Example I: Decentralized Matrix Completion

Consider a setting when the network of agents obtain incomplete observations of a matrix  $\theta_{\text{true}}$  of dimension  $m_1 \times m_2$  with  $m_1, m_2 \gg 0$ . The  $i$ th agent has corrupted observations from the *training* set  $\Omega_i \subset [m_1] \times [m_2]$  that are expressed as:

$$Y_{k,l} = [\theta_{\text{true}}]_{k,l} + Z_{k,l}, \quad \forall (k,l) \in \Omega_i. \quad (30)$$

To recover a low-rank  $\theta_{\text{true}}$ , we consider the following trace-norm constrained matrix completion (MC) problem:

$$\min_{\theta} \sum_{i=1}^N \sum_{(k,l) \in \Omega_i} f_i([\theta]_{k,l}, Y_{k,l}) \quad \text{s.t.} \quad \|\theta\|_{\sigma,1} \leq R, \quad (31)$$

where  $f_i : \mathbb{R}^2 \rightarrow \mathbb{R}$  is a loss function picked by agent  $i$  according to the observations he/she received.

Similar MC problems have been considered in [35–38], where [35] studied a consensus-based optimization based similar to ours and [36–38] studied the parallel computation setting where the agents are working synchronously in a fully connected network. Compared to our approach, these works assume that the rank of  $\theta_{\text{true}}$  is known in advance and solve the MC problem via matrix factorization. In addition, [35, 36] required that each local observation set  $\Omega_i$  to only have entries taken from a disjoint subset of the columns/rows only. Our approach does not require any prior knowledge or restrictions above.

We consider two different observation models. When  $Z_{k,l}$  is the i.i.d. Gaussian noise of variance  $\sigma_i^2$ , we choose  $f_i(\cdot, \cdot)$  to be the square loss function, i.e.,

$$f_i([\theta]_{k,l}, Y_{k,l}) := (1/\sigma_i^2) \cdot (Y_{k,l} - [\theta]_{k,l})^2. \quad (32)$$

This yields the classical MC problem in [5]. The next model considers the sparse+low rank matrix completion in [39], where the observations are contaminated with a sparse noise. Here, we model  $Z_{k,l}$  as a *sparse* noise in the sense that there are a few number of entries in  $\Omega_i$  where  $Z_{k,l}$  is non-zero. We choose  $f_i(\cdot, \cdot)$  to be the negated Gaussian loss, i.e.,

$$f_i([\boldsymbol{\theta}]_{k,l}, Y_{k,l}) := \left(1 - \exp\left(-\frac{([\boldsymbol{\theta}]_{k,l} - Y_{k,l})^2}{\sigma_i}\right)\right), \quad (33)$$

where  $\sigma_i > 0$  controls the robustness to outliers for the data obtained at the  $i$ th agent. Here,  $f_i(\cdot, \cdot)$  can be seen as a *smoothed*  $\ell_0$  loss [40] and gives enhanced robustness to outliers in the data. Notice that  $f_i(\cdot, \cdot)$  is non-convex and the resultant MC problem (31) is also non-convex.

Note that (31) is a special case of Problem (1) with  $\mathcal{C}$  being the trace-norm ball. The gossip-based DeFW algorithm can be applied on (31) directly. The projection-free nature of the DeFW algorithm leads to a low complexity implementation (31), especially when  $m_1, m_2 \gg 0$ . Lastly, the communication cost in the DeFW algorithm is analyzed below:

- The SAGA-like gradient surrogate  $\nabla_t^i F$  in (25) is supported only on  $\cup_{i=1}^N \Omega_i$  since for all  $i \in [N]$ ,

$$\nabla f_i(\bar{\boldsymbol{\theta}}_t^i) = \sum_{(k,l) \in \Omega_i} f'_i([\bar{\boldsymbol{\theta}}_t^i]_{k,l}, Y_{k,l}) \cdot \mathbf{e}_k(\mathbf{e}_l')^\top \quad (34)$$

is supported on  $\Omega_i$ , where  $\bar{\boldsymbol{\theta}}_t^i$  is defined in (21). In the above,  $\mathbf{e}_k$  ( $\mathbf{e}_l'$ ) is the  $k$ th ( $l$ th) canonical basis vector for  $\mathbb{R}^{m_1}$  ( $\mathbb{R}^{m_2}$ ) and  $f'_i(\theta, y)$  is the derivative of  $f_i(\theta, y)$  taken with respect to  $\theta$ . Consequently, the in-network average  $\overline{\nabla_t^i F}$  is supported only on  $\cup_{i=1}^N \Omega_i$ . As  $|\cup_{i=1}^N \Omega_i| \ll m_1 m_2$ , the amount of information exchanged during the *aggregating* step (Line 4 in DeFW) is low.

- The update direction  $\mathbf{a}_t^i$  is a rank-one matrix composed of the top singular vectors of  $\overline{\nabla_t^i F}$  (cf. (18)). Since every iteration in DeFW adds at most  $N$  distinct pair of singular vectors to  $\bar{\boldsymbol{\theta}}_t$ , the rank of  $\bar{\boldsymbol{\theta}}_t^i$  is upper bounded by  $t \cdot N$  if we initialize by  $\bar{\boldsymbol{\theta}}_0^i = \mathbf{0}$ . We can reduce the communication cost in the *consensus* step (Line 3 in DeFW) by exchanging these singular vectors. Note that we are exchanging  $(tN) \cdot (m_1 + m_2)$  entries instead of  $m_1 \cdot m_2$ .
- When the agents are *only* concerned with predicting the entries of  $\boldsymbol{\theta}_{\text{true}}$  in the subset  $\Xi \subset [m_1] \times [m_2]$ , instead of propagating the singular vectors as described above, the *consensus* step can be carried out by exchanging only the entries of  $\boldsymbol{\theta}_{t+1}^i$  in  $\Xi \cup (\cup_{i=1}^N \Omega_i)$  without affecting the operations of the DeFW algorithm. In this case, the communication cost is  $|\Xi \cup (\cup_{i=1}^N \Omega_i)|$ .

## 4.2 Example II: Communication Efficient DeFW for LASSO

Our next example considers applying the DeFW algorithm on a decentralized LASSO problem. Let  $(\mathbf{y}_i, \mathbf{A}_i)$  be the available data tuple at agent  $i \in [N]$  such that  $\mathbf{A}_i \in \mathbb{R}^{m \times d}$  and  $\mathbf{y}_i \in \mathbb{R}^m$ . The data  $\mathbf{y}_i$  is a corrupted measurement of some unknown parameter  $\boldsymbol{\theta}_{\text{true}}$ , following the model:

$$\mathbf{y}_i = \mathbf{A}_i \boldsymbol{\theta}_{\text{true}} + \mathbf{z}_i, \quad (35)$$

where  $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$  are independent noise vectors. Furthermore, we assume  $m \ll d$  such that the matrix  $\mathbf{A}_i^\top \mathbf{A}_i$  is rank-deficient. However, the parameter  $\boldsymbol{\theta}_{\text{true}}$  is  $s$ -sparse such that  $s = \|\boldsymbol{\theta}_{\text{true}}\|_0 \ll d$ .

This motivates us to consider the following distributed LASSO problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^N \frac{1}{2} \|\mathbf{y}_i - \mathbf{A}_i \boldsymbol{\theta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_1 \leq R, \quad (36)$$

Notice that the above is a special case of (1) with  $f_i(\boldsymbol{\theta}) = (1/2)\|\mathbf{y}_i - \mathbf{A}_i \boldsymbol{\theta}\|_2^2$  and  $\mathcal{C}$  is the  $\ell_1$ -ball in  $\mathbb{R}^d$ . We assume that (36) has an optimal solution  $\boldsymbol{\theta}^*$  that is sparse.

A number of state-of-the-art PGD-based decentralized algorithms are easily applicable to (36). For example, the D-PG algorithm in [10] can be described by the following simple recursion — at iteration  $t$ , we do

$$\boldsymbol{\theta}_{t+1}^{i,PG} = \mathcal{P}_{\mathcal{C}} \left( \sum_{j=1}^N W_{ij} \boldsymbol{\theta}_t^{j,PG} - \alpha_t \nabla f_i \left( \sum_{j=1}^N W_{ij} \boldsymbol{\theta}_t^{j,PG} \right) \right), \quad (37)$$

where  $\alpha_t \in (0, 1]$  is a diminishing step size and  $W_{ij}$  is the weight matrix described in the last section. As (36) is convex, the D-PG algorithm (37) is guaranteed to converge to an optimal solution  $\boldsymbol{\theta}^*$  of (36) at a rate of  $\mathcal{O}(1/t)$ .

During each iteration of the D-PG algorithm, the  $i$ th agent exchanges its current iterate  $\boldsymbol{\theta}_t^{i,PG}$  with the neighboring agents. Although  $\boldsymbol{\theta}^*$  is sparse, we notice that  $\boldsymbol{\theta}_t^{i,PG}$  computed in the D-PG recursion above is likely to be dense for any finite  $t$ . The per-iteration communication complexity/cost, i.e., defined as the number of non-zero real numbers exchanged per agent, for the D-PG algorithm is as high as  $\Theta(d)$ . This renders the D-PG algorithm unsuitable when the communication network between agents is limited in bandwidth. Despite the drawback of high communication complexity, we note that the (worst-case) per-iteration computation complexity of (37) is  $\mathcal{O}(d \log d)$  [30]. When  $d \gg 0$ , both the communication cost and computational complexity of D-PG may be high.

We notice that [6, 7] have considered distributed sparse recovery algorithm with focus on the communication efficiency. However, their algorithms are based on the iterative hard thresholding (IHT) formulation that requires a-priori knowledge on the sparsity level of  $\boldsymbol{\theta}_{\text{true}}$ .

The gossip-based DeFW algorithm in Section 3 can be applied directly to (36). However, we notice that similar issue as the D-PG algorithm may arise during the GAC update in the *aggregating* stage, since the gradient surrogate (25) is also dense. As a remedy, we propose a *modified* gossip-based DeFW algorithm for solving (36) in a *communication efficient* manner. The modified algorithm applies a ‘sparsification’ procedure to reduce the communication cost involved in the *aggregating step* and exploits the structure of the DeFW algorithm when the constraint set  $\mathcal{C}$  is an  $\ell_1$ -ball. Moreover, the flexible nature of our analysis allows us to analyze the performance of the modified DeFW algorithm.

In the modified DeFW algorithm, the *consensus step* in the gossip-based DeFW will be unchanged as this step has a low communication cost. Particularly, from (17) and (21), we see that  $\boldsymbol{\theta}_t^i$  is at most  $(t-1)N+1$ -sparse since  $\mathbf{a}_t^i$  is always a 1-sparse vector<sup>2</sup> (cf. (17)). As such, the communication cost of this step is always bounded by  $t \cdot N$ .

We then focus on the *aggregating step*. Observe that a naive implementation requires a communication cost of  $\Theta(d)$  as the gradient surrogate in (25) is dense in general. However, we recall from (17) that *only the largest magnitude coordinate* in  $\nabla_t^i F$  is sought when computing  $\mathbf{a}_t^i$  alone.

---

<sup>2</sup>As pointed out by [28], this observation also leads to an interesting sparsity-accuracy trade-off when applying FW on  $\ell_1$  constrained problems.

Therefore, as long as the largest magnitude coordinate in  $\overline{\nabla_t^i F}$  is preserved, the operation of the gossip-based DeFW algorithm will be unaffected. This motivates us to ‘sparsify’ the gradient information at each iteration before exchanging them with the neighboring agents. Let  $\Omega_t \subseteq [d]$  be the coordinate set whose gradient information is exchanged by the agents via the GAC protocol:

$$\widehat{\nabla_t^i F} := (\nabla f_i(\bar{\theta}_t^i)) \odot \mathbf{1}_{\Omega_t}, \quad \text{where } \mathbf{1}_{\Omega_t} = \sum_{k \in \Omega_t} \mathbf{e}_k, \quad (38)$$

and  $\odot$  denotes the Hadamard/element-wise product. Notice that a sparsified gradient (38) is used in the above in lieu of the SAGA-like gradient surrogate (25). This is necessary as the selected coordinate set  $\Omega_t$  changes with the iteration number and the *full* average gradient  $\overline{\nabla_{t-1}^i F}$  from the previous iteration is unavailable.

Let  $\ell_t = \lceil C_l + \log(t) / \log |\lambda_2^{-1}(\mathbf{W})| \rceil$  where  $C_l$  is some finite constant and  $\lambda_2(\mathbf{W})$  is the second largest eigenvalue of the weight matrix  $\mathbf{W}$ , the modified gossip-based DeFW algorithm computes the approximate gradient average  $\overline{\nabla_t^i F}$  in line 4 of Algorithm 1 by:

$$\overline{\nabla_t^i F} = \sum_{j=1}^N [\mathbf{W}^{\ell_t}]_{ij} \cdot \widehat{\nabla_t^j F}. \quad (39)$$

It is important to remark that (39) requires  $\ell_t$  rounds of GAC updates to be performed at iteration  $t$ , i.e., a logarithmically increasing number of rounds of GAC updates. The update directions  $\mathbf{a}_t^i$  can then be computed using the  $\overline{\nabla_t^i F}$  computed in (39). Furthermore, as  $\overline{\nabla_t^i F}$  is at most  $|\Omega_t|$ -sparse, solving the LO (7b) requires a worst-case complexity of  $\mathcal{O}(|\Omega_t|)$ .

We pick our coordinate set  $\Omega_t$  in a decentralized manner. Consider the following decomposition:

$$\Omega_t = \bigcup_{i=1}^N \Omega_{t,i}, \quad (40)$$

where  $\Omega_{t,i} \subset [d]$  is picked by agent  $i$  at iteration  $t$ . The coordinate set  $\Omega_t$  needs to be known by all agents before (39), this can be achieved with a low communication overhead, e.g., by forming a spanning tree on the graph  $G$  and broadcasting the required indices in  $\Omega_t$  to all agents; see [41]. Set  $p_t$  as the maximum desirable cardinality of  $\Omega_{t,i}$ , agent  $i$  chooses the coordinate set using one of the following two schemes:

- *Random Co-ord. (rand)* — each agent chooses  $\Omega_{t,i}$  by selecting  $p_t$  coordinates from  $[d]$  by sampling uniformly with replacement.
- *Extreme Co-ord. (extreme)* — each agent chooses  $\Omega_{t,i}$  as the  $p_t$  largest magnitude coordinates of the local gradient vector  $\nabla f_i(\bar{\theta}_t^i)$ .

For the random coordinate selection scheme, the following lemma shows that the gradient approximation error can be controlled at a desirable rate with an appropriate choice of  $p_t$ . Let  $\xi_t := (1 - (1 - 1/d)^{p_t N})$ , we have:

**Lemma 3** *Set  $\epsilon > 0$  and  $\ell_t = \lceil C_l + \log(t) / \log |\lambda_2^{-1}(\mathbf{W})| \rceil$ . Let  $p_t = \Omega(t)$ . With probability at least  $1 - \pi^2 \epsilon / 6$ , the following holds for all  $\theta \in \mathcal{C}$ :*

$$\left\| \xi_t^{-1} \overline{\nabla_t^i F} - \frac{1}{N} \sum_{i=1}^N \nabla f_i(\bar{\theta}_t^i) \right\|_{\infty} = \mathcal{O}\left(\frac{\sqrt{\log(t^2/\epsilon)}}{t}\right), \quad (41)$$

for all  $t \geq 1$  and  $i \in [N]$ .

The proof can be found in Appendix E.

It is important to note that the above result is given in terms of  $\xi_t^{-1}\overline{\nabla_t^i F}$  instead of  $\overline{\nabla_t^i F}$ . However, the result remains relevant as the LO subproblem (7b) in the FW step of the DeFW algorithm is *scale invariant*, i.e.,  $\arg \min_{\mathbf{a} \in \mathcal{C}} \langle \overline{\nabla_t^i F}, \mathbf{a} \rangle = \arg \min_{\mathbf{a} \in \mathcal{C}} \langle \alpha \overline{\nabla_t^i F}, \mathbf{a} \rangle$  for any  $\alpha > 0$ . In other words, performing the FW step with  $\overline{\nabla_t^i F}$  is the same as doing so with  $\xi_t^{-1}\overline{\nabla_t^i F}$ . As  $\xi_t^{-1}\overline{\nabla_t^i F}$  is an  $\mathcal{O}(1/t)$  approximation to  $N^{-1} \sum_{j=1}^N \nabla f_j(\bar{\boldsymbol{\theta}}_t^j)$ , H2 is satisfied with  $\Delta d_t = \mathcal{O}(1/t)$  and the convergence results in Theorem 1 and 2 apply.

The analysis for extreme coordinate selection scheme depends on the statistics of the gradient vector  $\nabla f_i(\bar{\boldsymbol{\theta}}_t^i)$  and does not lead to an insightful expression without imposing further assumptions. Nevertheless, intuitively the scheme may perform better than the uniform randomized scheme as the maximum magnitude coordinate of  $N^{-1} \sum_{i=1}^N \nabla f_i(\bar{\boldsymbol{\theta}}_t^i)$  is likely to be included in  $\Omega_t$ . This intuition will be confirmed by our numerical examples. Finally, we conclude that

**Corollary 2** *The modified gossip-based DeFW algorithm using **rand** coordinate selection, i.e., with line 4 & 3 in Algorithm 1 replaced by (39) & (21), respectively, generates iterates that satisfy the guarantees in Theorem 1 (with high probability). Under strong convexity and interior optimal point assumption, the communication complexity is  $\Theta(N \cdot (1/\delta) \cdot \log(1/\delta))$  to reach a  $\delta$ -optimal solution to (36).*

The first statement above is a consequence of Lemma 3. The second statement can be verified by noting that reaching a  $\delta$ -optimal solution requires  $\Theta(1/\sqrt{\delta})$  iterations and the communication cost is  $\Theta(Nt \log t)$  at iteration  $t$ , as the agents exchange an  $\Theta(t \cdot N)$ -sparse vector for  $\Theta(\log t)$  times.

## 5 Numerical Experiments

We perform numerical experiments to verify our theoretical findings on the DeFW algorithm. The following discussions will focus on the two applications described in the previous section and are verified using synthetic and real datasets. In this section, to simulate the decentralized optimization setting, we artificially construct a network of  $N = 50$  agents, where the underlying communication network  $G$  is an Erdos-Renyi graph with connectivity of 0.1. For the GAC steps (21), (26) & (39), the doubly stochastic matrix  $\mathbf{W}$  is calculated according to the Metropolis-Hastings rule in [42].

### 5.1 Decentralized Matrix Completion

This section considers the decentralized matrix completion problem, where the goal is to predict the missing entries of an unknown matrix through corrupted partial measurements.

Consider two datasets — the first dataset is synthetically generated where the unknown matrix  $\boldsymbol{\theta}_{\text{true}}$  is rank- $K$  and it has a dimension of  $m_1 = 100$  and  $m_2 = 250$ ; the matrix is  $\boldsymbol{\theta}_{\text{true}} = \sum_{i=1}^K \mathbf{u}_i \mathbf{v}_i^\top / K$  where  $\mathbf{u}_i, \mathbf{v}_i$  have i.i.d.  $\mathcal{N}(0, 1)$  entries and different settings of  $K$  will be experimented. The second dataset is taken from the `movielens` repository [43], where we consider the `movielens100k` dataset. Here, the unknown matrix  $\boldsymbol{\theta}_{\text{true}}$  records the movie ratings of  $m_1 = 943$  users on  $m_2 = 1682$  movies; and a total of  $10^5$  entries in  $\boldsymbol{\theta}_{\text{true}}$  are available as integers ranging from 1 to 5. We divide the dataset into a training and testing sets and evaluate the mean square error (MSE) on the testing set as:

$$\text{MSE} = |\Omega_{\text{test}}|^{-1} \sum_{(k,l) \in \Omega_{\text{test}}} |[\boldsymbol{\theta}_{\text{true}}]_{k,l} - [\hat{\boldsymbol{\theta}}]_{k,l}|^2, \quad (42)$$

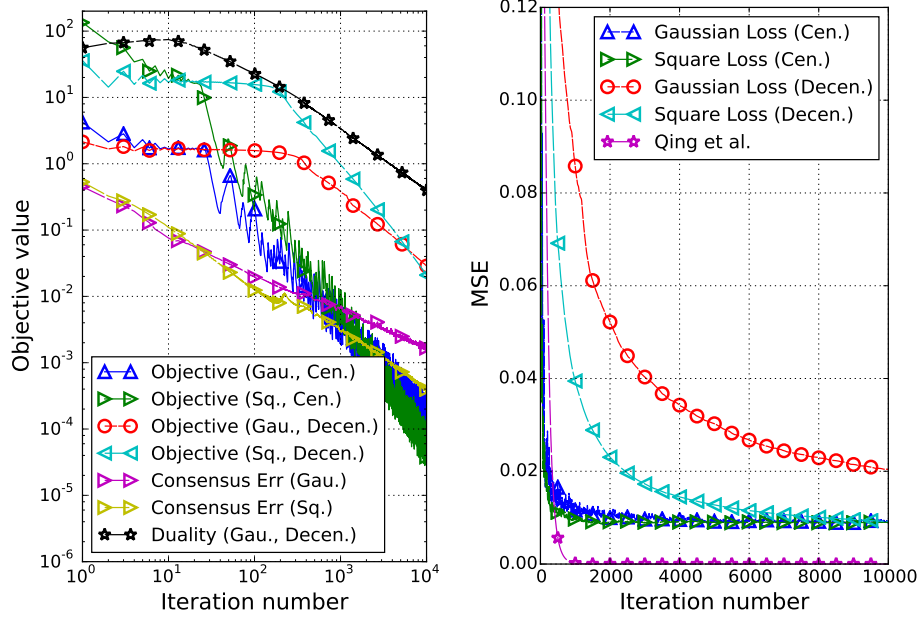


Figure 1: Convergence on noiseless synthetic data with rank  $K = 5$ . (Left) Objective values and consensus error of  $\bar{\theta}_t^i$  against the DeFW iteration number  $t$ , the objective values are evaluated by  $F(\bar{\theta}_t)$ . (Right) Worst-case MSE (among agents) against iteration number on the testing set. The legend ‘Gau.’, ‘Sq.’ denote the gossip-based DeFW algorithm applied to (31) with the negated Gaussian and square loss, respectively.

where  $\hat{\theta}$  denotes the estimated  $\theta$  produced by the algorithm.

For the synthetic dataset, the training (testing) set contains  $0.5 \times 10^4$  ( $2 \times 10^4$ ) entries which are selected randomly. For *movielens100k*, the training (testing) set contains  $80 \times 10^3$  ( $20 \times 10^3$ ) entries. The training data of the two datasets are equally partitioned into  $N = 50$  parts, i.e., for the synthetic dataset, each agent holds 100 entries; for *movielens100k*, each agent holds 1600 entries. We compare the performance of the proposed gossip-based DeFW algorithm with square loss and negated Gaussian loss, as described in Section 4.1. Unless otherwise specified, we fix the number of GAC rounds applied at  $\ell = 1$  such that the agents only exchange information once per iteration. As the negated Gaussian loss is non-convex, we set the step size as  $\gamma_t = 1/t^{0.75}$ . The centralized FW algorithm for both losses will also be compared (cf. (7)); as well as the consensus based algorithm proposed in [35] (labeled as ‘Qing et al.’ in the legends).

Our first example considers the noiseless synthetic dataset with  $K = 5$  and the results are shown in Fig. 1. Here, for the DeFW algorithms, we set the trace-norm radius to  $R = 1.2\|\theta_{\text{true}}\|_{\sigma,1}$ ; and the algorithm in [35] is supplied with the true rank  $K$  of  $\theta_{\text{true}}$ . Notice that for this set of data, the minimum of (31) can be achieved by  $\theta = \theta_{\text{true}} \in \mathcal{C}$  with a zero optimal objective value. From the left figure, for the DeFW algorithm applied to the convex square loss function, we observe an  $\mathcal{O}(1/t^2)$  trend for the objective values, corroborating with our analysis in Theorem 1; for the non-convex Gaussian loss function, the objective value and the duality gap  $g_t = \max_{\theta \in \mathcal{C}} \langle \nabla F(\bar{\theta}_t), \bar{\theta}_t - \theta \rangle$  in Theorem 2 also decays with  $t$ , indicating the convergence to a stationary point. Moreover, the consensus error of  $\bar{\theta}_t^i$  for DeFW applied to the two objective functions decay at the rate predicted by Lemma 1. On the other hand, the right figure compares mean square error (MSE) of the predicted

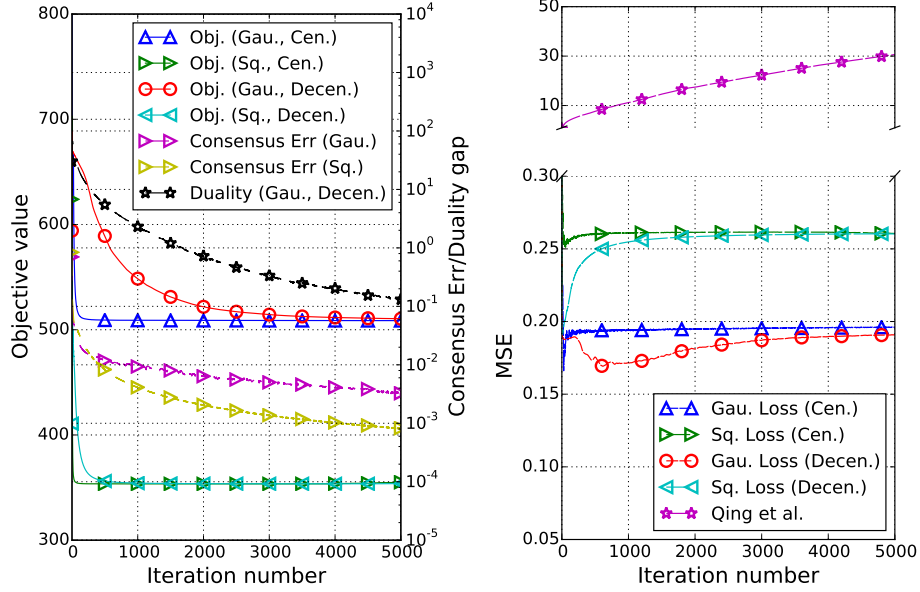


Figure 2: Convergence on sparse-noise contaminated synthetic data with rank  $K = 5$ . (Left) Objective values and consensus error of  $\hat{\theta}_t^i$  against the DeFW iteration number  $t$ . Notice that the consensus error (in purple and yellow) / duality gaps (in black) are plotted in a logarithmic scale (cf. the right y-axis) while the objective values are plotted in a linear scale; (Right) MSE against the DeFW iteration number  $t$  on the testing set.

matrix  $\theta$  for the testing set. Here, we also compare the result with the decentralized MC algorithm in [35]. We observe that the MSE performance of the DeFW algorithms approach their centralized counterpart as the iteration number grows, yet the algorithm in [35] achieves the best performance in this setting, notice that the true rank of  $\theta_{\text{true}}$  is provided to this algorithm. However, as we shall see next, the algorithm in [35] may not be robust enough when the observations are contaminated with noise or the rank of  $\theta_{\text{true}}$  increases.

The second example considers adding noise to the observations. In particular, we adopt the same setting as the previous example but include a *sparse* noise in the observations — here, each  $Z_{k,l} = p_{k,l} \cdot \tilde{Z}_{k,l}$  where  $p_{k,l}$  is Bernoulli with  $P(p_{k,l} = 1) = 0.2$  and  $\tilde{Z}_{k,l} \sim \mathcal{N}(0, 5)$  (cf. (30)). The convergence results are compared in Fig. 2. For the left figure, we observe similar convergence behaviors for the DeFW algorithms applied to different objective functions as in the previous example. On the right figure, we observe that the DeFW algorithm based on the Gaussian loss achieves the best MSE performance, demonstrating its robustness to outlier noise. We also see that the algorithm in [35] performs poorly on this dataset, i.e., the MSE achieved is greater than the MSE achieved by DeFW by an order of magnitude.

Another interesting discovery is that the algorithm in [35] seems to fail when the rank of  $\theta_{\text{true}}$  is high, even when the true rank is known and the observations are noiseless. In Fig. 3, we show the MSE against iteration number of the algorithms when the synthetic data is noiseless and generated with  $K = 10$ . As seen, the algorithm in [35] fails to produce a low MSE, while the DeFW algorithm offers a reasonable performance. We suspect that this is due to the difference in the optimization problem formulation, as a trace-norm constrained formulation (31) is tackled by the DeFW algorithm.

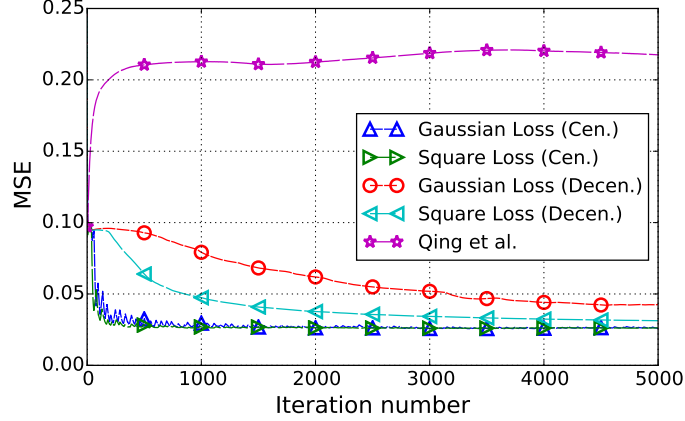


Figure 3: Convergence of test MSE against iteration number on the testing set on noise-free synthetic data with rank  $K = 10$ .

Lastly, we consider the dataset `movielens100k`. We set  $R = 10^5$  and focus on the MSE evaluated on the testing set against the iteration number for the proposed DeFW algorithm applied to different loss functions. The numerical results are presented in Fig. 4, where we also compare the case when we apply multiple ( $\ell = 1, 3$ ) rounds of GAC updates per iteration to speed up the algorithm. The left figure considers the noiseless scenario. As seen, the proposed DeFW algorithm applied on different loss functions converge to a reasonable MSE that is attained by the centralized FW algorithm. We see that the DeFW with Gaussian loss has a slower convergence compared to the square loss which is possibly attributed to the non-convexity of the loss function. Moreover, the algorithms achieve much faster convergence if we allow  $\ell = 3$  GAC rounds of network information exchange per iteration. On the other hand, the right figure considers when the observations are contaminated with a sparse noise following the same model in Fig. 2. Here, we observe that the Gaussian loss implementations attain the best MSE as the non-convex loss is more robust against the outliers in the training data. Interestingly, the DeFW algorithm with  $\ell = 3$  GAC rounds has even outperformed its centralized counterpart. We suspect that this is caused by the DeFW algorithm converging to a different local minima in the non-convex problem.

We remark that the DeFW algorithm applied to (31) has a noticeable computational advantage over the PGD-based algorithm such as D-PG [10]. Taking the `movielens100k` dataset as an example (with  $m_1 = 943, m_2 = 1682$ ), the FW step in line 5 of the DeFW algorithm takes about  $\sim 3 \times 10^{-2}$  seconds<sup>3</sup> to complete on each agent since only the principal singular vectors are required. On the other hand, each iteration of D-PG requires computing a full SVD for a dense matrix, which takes  $\sim 0.8$  seconds to complete on each agent for the considered problem size. The complexity gap is further widened when we consider larger problems.

## 5.2 Communication-efficient Sparse Learning

This section conducts numerical experiments on the decentralized sparse learning problem. We focus on the *modified DeFW algorithm* in Section 4.2 that has better communication efficiency. We evaluate the performance of DeFW on both synthetic data and benchmark dataset. For the synthetic data, we randomly generate each  $\mathbf{A}_i$  as a  $(m = 20) \times (d = 10000)$  matrix with  $\mathcal{N}(0, 1)$

<sup>3</sup>Tested on a quad-core Intel Core i7 laptop running MATLAB 2015b.



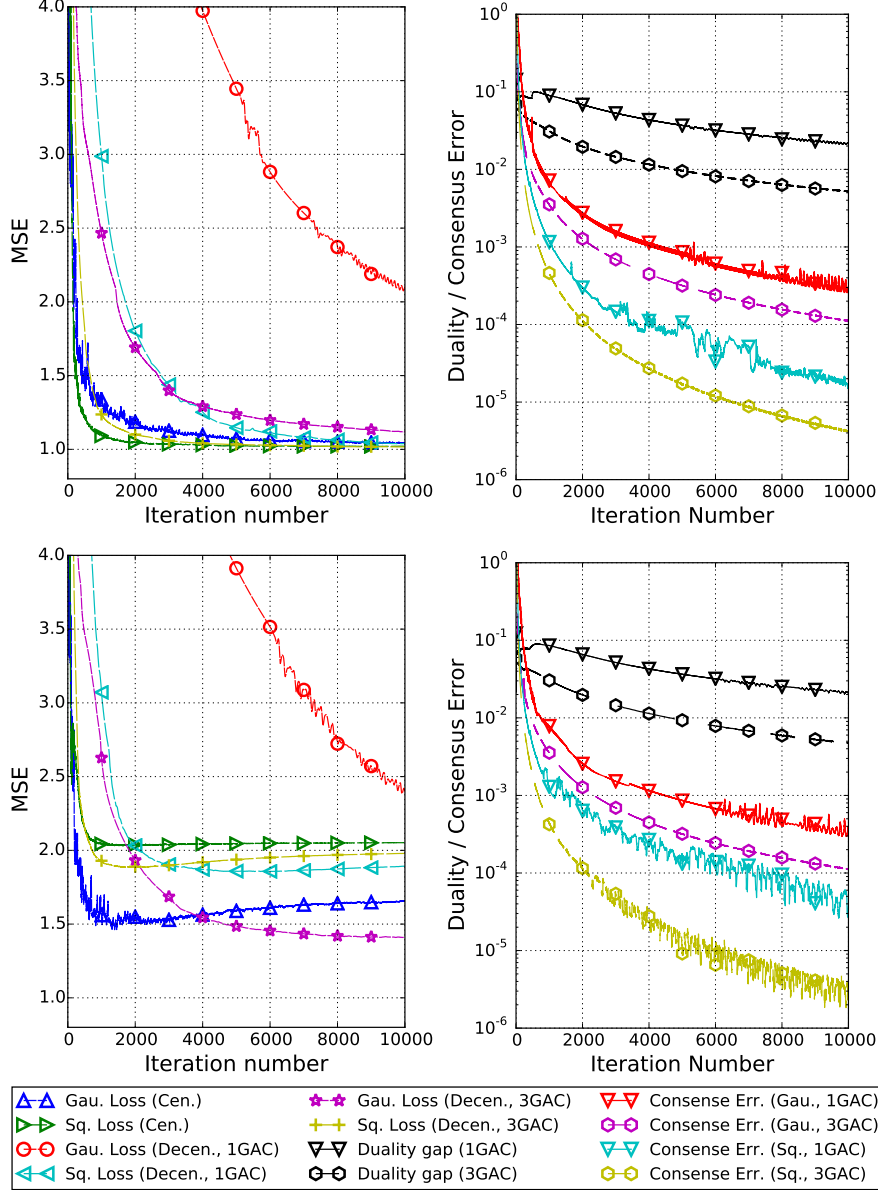


Figure 4: Convergence of the DeFW algorithm on *movielens100k* dataset with different loss functions. (Top) Noiseless observations; (Bottom) sparse-noise contaminated observations. Note that the duality gap / consensus errors are plotted in a logarithmic scale in the right figures.

elements (cf. (35)) and  $\theta_{\text{true}}$  is a random sparse vector with  $\|\theta_{\text{true}}\|_0 = 50$  such that the non-zero elements are also  $\mathcal{N}(0,1)$ . The observation noise  $z_i$  has a variance of  $\sigma^2 = 0.01$ . We also apply the DeFW algorithm on *sparco7* [44], which is a commonly used dataset for benchmarking sparse recovery algorithms. For *sparco7*, we have  $A_i \in \mathbb{R}^{12 \times 2560}$  as the local measurement matrix and  $\theta_{\text{true}}$  is a sparse vector with  $\|\theta_{\text{true}}\|_0 = 20$ .

The modified DeFW algorithm is implemented with  $p_t = \lceil 2 + \alpha_{\text{comm}} \cdot t \rceil$ ,  $\ell_t = \lceil \log(t) + 1 \rceil$  with extreme or random coordinate selection. We compare the algorithms of PG-EXTRA [11],

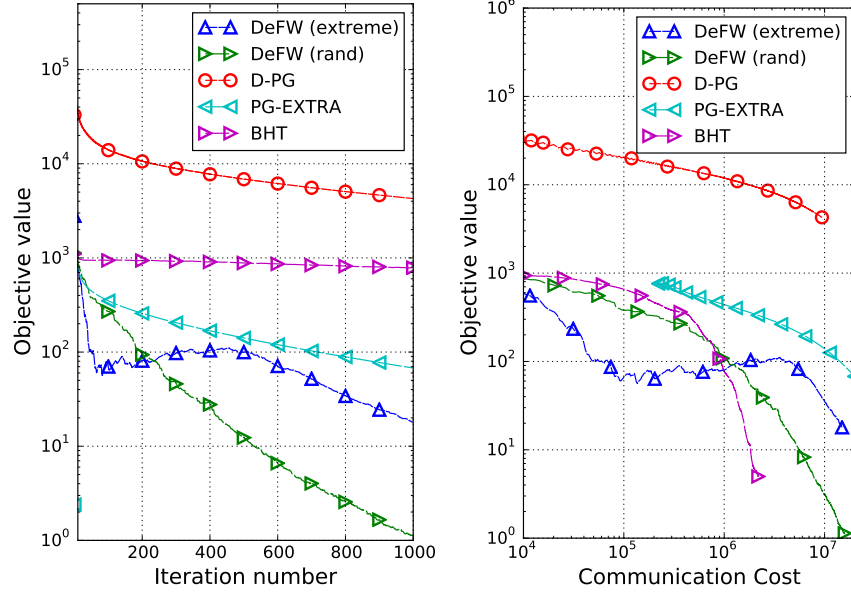


Figure 5: Convergence of the objective value on LASSO with synthetic dataset. (Left) against the iteration number. (Right) against the communication cost (i.e., total number of values transmitted/received in the network during GAC updates). In the legend, ‘DeFW (extreme)’ refers to the extreme coordinate selection and ‘DeFW (rand)’ refers to the random coordinate selection scheme.

D-PG [10] and BHT [6]. DeFW, PG-EXTRA and D-PG are set to solve the convex problem (36) with  $R = 1.1\|\theta_{\text{true}}\|_1$ . BHT is a communication efficient distributed implementation of IHT and is supplied with the true sparsity level in our simulations.

The first example in Fig. 5 shows the convergence of the algorithms on the synthetic dataset, where we compare the objective value against the number of iterations and the communication cost, i.e., total number of values sent during the distributed optimization. We set  $\alpha_{\text{comm}} = 0.05$  for the DeFW algorithms. From the left figure, we observe that DeFW and PG-EXTRA algorithms have similar iteration complexity while ‘DeFW (rand)’ seems to have the fastest convergence. Meanwhile, BHT requires a high number of iterations for convergence. On the other hand, in the right figure, the DeFW algorithms demonstrate the best communication efficiency at low accuracy, while they lose to BHT at higher accuracy. We speculate that this is due to the geometric convergence rate offered by the IHT algorithm [45]. Moreover, ‘DeFW (extreme)’ achieves a better accuracy at the beginning of the iterations (i.e., less communication cost paid) but is overtaken by ‘DeFW (rand)’ as the communication cost grows.

We then compare the performance on **sparco7**, where we show the convergence of objective value against the communication cost in Fig. 6. We set  $\alpha_{\text{comm}} = 0.025$  for the DeFW algorithms. From the figure, we observe that at low accuracy, the DeFW algorithms offer the best communication cost-accuracy trade-off, i.e., it performs the best at an accuracy of above  $\sim 10^{-2}$ . Moreover, ‘DeFW (extreme)’ seems to perform better than ‘DeFW (rand)’ in this example. Nevertheless, the BHT algorithm achieves the best performance when the communication cost paid is above  $3 \times 10^5$ .

Lastly, we comment that although BHT seems to require the lowest communication cost at the *high* accuracy regime, its computational complexity is generally high as BHT requires a much larger number of iterations to reach a reasonable accuracy (as evidenced in Fig. 5 (Left)). In light

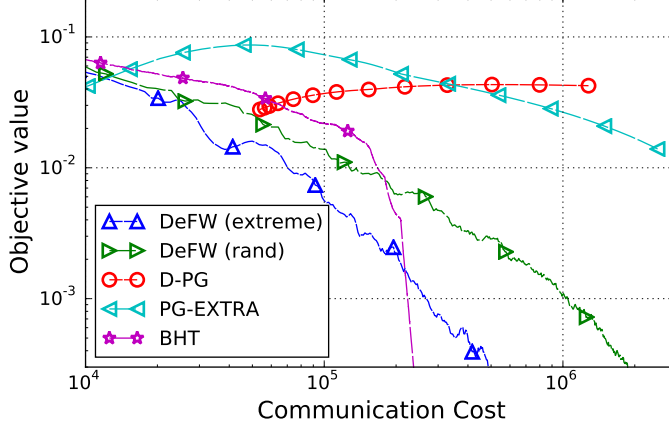


Figure 6: Convergence of the objective value against the communication cost on LASSO with `sparco7` dataset. In the legend, ‘DeFW (extreme)’ refers to the extreme coordinate selection and ‘DeFW (rand)’ refers to the random coordinate selection scheme.

of this, the modified DeFW offers a better balance between the communication and computation complexity.

## 6 Conclusions & Open Problems

In this paper, we have studied a decentralized projection-free algorithm for constrained optimization, which we called the DeFW algorithm. Importantly, we showed that the DeFW algorithm converges for both convex and non-convex loss functions and the respective convergence rates are analyzed. The efficacy of the proposed algorithm is demonstrated through tackling two problems related to machine learning and signal processing, with the advantages over previous state-of-the-art demonstrated through numerical experiments. Future directions of study on projection-free algorithm for decentralized optimization include an asynchronous version of the DeFW algorithm.

## A Proof of Theorem 1

The proof of Theorem 1 follows from our recent analysis on online/stochastic FW algorithm [46]. Using line 5 of Algorithm 1, we observe that:

$$\bar{\theta}_{t+1} = \bar{\theta}_t + \gamma_t (N^{-1} \sum_{i=1}^N \mathbf{a}_t^i - \bar{\theta}_t). \quad (43)$$

Define  $h_t := F(\bar{\theta}_t) - F(\theta^*)$  where  $\theta^*$  is an optimal solution to (1). From the  $L$ -smoothness of  $F$  and the boundedness of  $\mathcal{C}$ , we have:

$$h_{t+1} \leq h_t + \frac{\gamma_t}{N} \sum_{i=1}^N \langle \mathbf{a}_t^i - \bar{\theta}_t, \nabla F(\bar{\theta}_t) \rangle + \gamma_t^2 \frac{L\bar{\rho}^2}{2}, \quad (44)$$

where  $\bar{\rho}$  was defined in (5). Observe the following chain for the inner product: for each  $i \in [N]$ , we have

$$\begin{aligned} \langle \mathbf{a}_t^i - \bar{\boldsymbol{\theta}}_t, \nabla F(\bar{\boldsymbol{\theta}}_t) \rangle &\leq \langle \mathbf{a}_t^i - \bar{\boldsymbol{\theta}}_t, \overline{\nabla_t^i F} \rangle + \bar{\rho} \|\overline{\nabla_t^i F} - \nabla F(\bar{\boldsymbol{\theta}}_t)\|_2 \\ &\leq \langle \mathbf{a} - \bar{\boldsymbol{\theta}}_t, \overline{\nabla_t^i F} \rangle + \bar{\rho} \cdot \|\overline{\nabla_t^i F} - \nabla F(\bar{\boldsymbol{\theta}}_t)\|_2, \quad \forall \mathbf{a} \in \mathcal{C} \\ &\leq \langle \mathbf{a} - \bar{\boldsymbol{\theta}}_t, \nabla F(\bar{\boldsymbol{\theta}}_t) \rangle + 2\bar{\rho} \cdot \|\overline{\nabla_t^i F} - \nabla F(\bar{\boldsymbol{\theta}}_t)\|_2, \quad \forall \mathbf{a} \in \mathcal{C}, \end{aligned} \quad (45)$$

where we have added and subtracted  $\overline{\nabla_t^i F}$  in the first inequality; and used the fact  $\mathbf{a}_t^i = \arg \min_{\mathbf{a} \in \mathcal{C}} \langle \mathbf{a}, \overline{\nabla_t^i F} \rangle$  in the second inequality. Recalling that  $\overline{\nabla_t F} = N^{-1} \sum_{i=1}^N \nabla f_i(\bar{\boldsymbol{\theta}}_t^i)$ , we observe

$$\begin{aligned} \|\overline{\nabla_t^i F} - \nabla F(\bar{\boldsymbol{\theta}}_t)\|_2 &\leq \|\overline{\nabla_t^i F} - \overline{\nabla_t F}\|_2 + \|\overline{\nabla_t F} - \nabla F(\bar{\boldsymbol{\theta}}_t)\|_2 \\ &\leq \Delta d_t + N^{-1} \sum_{i=1}^N \|\nabla f_i(\bar{\boldsymbol{\theta}}_t^i) - \nabla f_i(\bar{\boldsymbol{\theta}}_t)\|_2 \\ &\leq \Delta d_t + L \cdot N^{-1} \sum_{i=1}^N \|\bar{\boldsymbol{\theta}}_t^i - \bar{\boldsymbol{\theta}}_t\|_2 \\ &\leq \Delta d_t + L \cdot \Delta p_t, \end{aligned} \quad (46)$$

where the third inequality is due to the  $L$ -smoothness of  $\{f_i\}_{i=1}^N$ . Recalling that  $\Delta p_t = C_p/t$ ,  $\Delta d_t = C_g/t$  and substituting the results above into the inequality (44) implies:

$$h_{t+1} \leq h_t + \gamma_t \langle \bar{\mathbf{a}}_t - \bar{\boldsymbol{\theta}}_t, \nabla F(\bar{\boldsymbol{\theta}}_t) \rangle + \gamma_t^2 \frac{L\bar{\rho}^2}{2} + 2\bar{\rho}\gamma_t \frac{C_g + LC_p}{t}, \quad (47)$$

for some  $C < \infty$  that depends on  $L$  and the scaling constants of  $\Delta p_t, \Delta d_t$ , where  $\bar{\mathbf{a}}_t \in \mathcal{C}$  is the minimizer of the linear optimization (7b) using  $\nabla F(\bar{\boldsymbol{\theta}}_t)$ , i.e.,

$$\bar{\mathbf{a}}_t = \arg \min_{\mathbf{a} \in \mathcal{C}} \langle \mathbf{a}, \nabla F(\bar{\boldsymbol{\theta}}_t) \rangle. \quad (48)$$

**Case 1:** When  $F$  is convex, we observe

$$\langle \bar{\mathbf{a}}_t - \bar{\boldsymbol{\theta}}_t, \nabla F(\bar{\boldsymbol{\theta}}_t) \rangle \leq \langle \boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}_t, \nabla F(\bar{\boldsymbol{\theta}}_t) \rangle \leq -h_t, \quad (49)$$

where the first inequality is due to the optimality of  $\bar{\mathbf{a}}_t$  and the last inequality stems from the convexity of  $F$ . Plugging the above into (47) yields

$$h_{t+1} \leq (1 - \gamma_t)h_t + \gamma_t^2 \frac{L\bar{\rho}^2}{2} + \gamma_t \frac{2\bar{\rho}(C_g + LC_p)}{t}. \quad (50)$$

As  $\gamma_t = 2/(t+1)$ , from a high-level point of view, the above inequality behaves similarly to  $h_{t+1} \leq (1 - (1/t))h_t + \mathcal{O}(1/t^2)$ . Consequently, applying [47, Lemma 4] yields a  $\mathcal{O}(1/t)$  convergence rate for  $h_t$ . In fact, this is a deterministic version of the case analyzed by [46, Theorem 10]. In particular, setting  $\alpha = 1, K = 2$  in [46, (56)] and using an induction argument yield

$$h_t \leq 2 \cdot (4\bar{\rho}(C_g + LC_p) + L\bar{\rho}^2)/(t+1), \quad \forall t \geq 1. \quad (51)$$

**Case 2:** For the case when  $F$  is  $\mu$ -strongly convex and  $\boldsymbol{\theta}^*$  lies in the interior of  $\mathcal{C}$  with distance  $\delta > 0$  (cf. (6)). Using [46, Lemma 6], we have

$$\langle \bar{\boldsymbol{\theta}}_t - \bar{\mathbf{a}}_t, \nabla F(\bar{\boldsymbol{\theta}}_t) \rangle \geq \sqrt{2\mu\delta^2 h_t}. \quad (52)$$

Plugging the above into (47) gives

$$h_{t+1} \leq \sqrt{h_t}(\sqrt{h_t} - \gamma_t \sqrt{2\mu\delta^2}) + \gamma_t^2 \frac{L\bar{\rho}^2}{2} + \gamma_t \frac{2\bar{\rho}(C_g + LC_p)}{t}. \quad (53)$$

Compared to the case analyzed in (50), when  $h_t$  is decreased, the decrement in  $h_{t+1}$  is increased, leading to a faster convergence. This is a deterministic version of the case analyzed in [46, Theorem 7]. Setting  $\alpha = 1, K = 2$  in [46, (48)] and using an induction argument yields

$$h_t \leq \frac{(4\bar{\rho}(C_g + LC_p) + L\bar{\rho}^2)^2}{2\delta^2\mu} \cdot \frac{9}{(t+1)^2}, \quad \forall t \geq 1. \quad (54)$$

## B Proof of Theorem 2

### B.1 Convergence rate

To facilitate our following discussions, let us define the following *FW gap*:

$$g_t := \max_{\bar{\theta} \in \mathcal{C}} \langle \nabla F(\bar{\theta}_t), \bar{\theta}_t - \theta \rangle = \langle \nabla F(\bar{\theta}_t), \bar{\theta}_t - \bar{a}_t \rangle, \quad (55)$$

where the last equality is due to (48) from the previous proof. For simplicity, we shall assume that  $T$  is even in the following.

From the  $L$ -smoothness of  $F$ , we have:

$$F(\bar{\theta}_{t+1}) \leq F(\bar{\theta}_t) + \langle \nabla F(\bar{\theta}_t), \bar{\theta}_{t+1} - \bar{\theta}_t \rangle + \frac{L}{2} \|\bar{\theta}_{t+1} - \bar{\theta}_t\|_2^2. \quad (56)$$

We observe that:

$$\bar{\theta}_{t+1} - \bar{\theta}_t = N^{-1} \sum_{i=1}^N \gamma_t (\mathbf{a}_t^i - \bar{\theta}_t^i). \quad (57)$$

As  $\mathbf{a}_t^i, \bar{\theta}_t^i \in \mathcal{C}$ , we have  $\|\bar{\theta}_{t+1} - \bar{\theta}_t\|_2 \leq \gamma_t \bar{\rho}$ . Using (45) and (46) from the previous proof of Theorem 1, the inequality (56) can be bounded as:

$$\begin{aligned} F(\bar{\theta}_{t+1}) &\leq F(\bar{\theta}_t) - \gamma_t \langle \nabla F(\bar{\theta}_t), \bar{\theta}_t - \bar{a}_t \rangle \\ &\quad + 2\gamma_t \bar{\rho} \cdot (\Delta d_t + L \cdot \Delta p_t) + \gamma_t^2 L \bar{\rho}^2 / 2 \\ &= F(\bar{\theta}_t) - \gamma_t g_t + 2\gamma_t \bar{\rho} \cdot (\Delta d_t + L \cdot \Delta p_t) + \gamma_t^2 \frac{L \bar{\rho}^2}{2}. \end{aligned} \quad (58)$$

From the definition, we observe that  $g_t \geq 0$ . Now, summing the two sides of (58) from  $t = T/2 + 1$  to  $t = T$  gives:

$$\sum_{t=T/2+1}^T \gamma_t g_t \leq \sum_{t=T/2+1}^T \left( F(\bar{\theta}_t) - F(\bar{\theta}_{t+1}) \right) + \sum_{t=T/2+1}^T \left( 2\gamma_t \bar{\rho} \cdot (\Delta d_t + L \cdot \Delta p_t) + \gamma_t^2 \frac{L \bar{\rho}^2}{2} \right). \quad (59)$$

Canceling duplicated terms in the first term of the right hand side above gives:

$$\sum_{t=T/2+1}^T \gamma_t g_t \leq F(\bar{\theta}_{T/2+1}) - F(\bar{\theta}_{T+1}) + \sum_{t=T/2+1}^T \left( 2\gamma_t \bar{\rho} \cdot (\Delta d_t + L \cdot \Delta p_t) + \gamma_t^2 \frac{L \bar{\rho}^2}{2} \right). \quad (60)$$

As  $g_t, \gamma_t \geq 0$ , we can lower bound the left hand side as:

$$\sum_{t=T/2+1}^T \gamma_t g_t \geq \left( \min_{t \in [T/2+1, T]} g_t \right) \cdot \left( \sum_{t=T/2+1}^T \gamma_t \right), \quad (61)$$

and observe that for all  $T \geq 6$  and  $\alpha \in (0, 1)$ ,

$$\sum_{t=T/2+1}^T \gamma_t \geq \frac{T^{1-\alpha}}{1-\alpha} \cdot \left( 1 - \left( \frac{2}{3} \right)^{1-\alpha} \right) = \Omega(T^{1-\alpha}). \quad (62)$$

Define the constant  $C := L\bar{\rho}^2/2 + 2\bar{\rho}(C_g + LC_p)$ . When  $\alpha \geq 0.5$ , using the fact that  $\gamma_t = t^{-\alpha}$ ,  $\Delta p_t = C_p/t^\alpha$ ,  $\Delta d_t = C_g/t^\alpha$ , the right hand side of (60) is bounded above by:

$$G \cdot \rho + C \cdot \sum_{t=T/2+1}^T t^{-2\alpha} \leq G \cdot \rho + C \cdot \log 2, \quad (63)$$

note that the series is converging as we are summing from  $t = T/2 + 1$  to  $t = T$ . Dividing the above term by the lower bound (62) to  $\sum_{t=T/2+1}^T \gamma_t$  yields (15).

On the other hand, when  $\alpha < 0.5$ , we notice that

$$\sum_{t=T/2+1}^T t^{-2\alpha} \leq \int_{T/2}^T t^{-2\alpha} dt = \frac{2^{1-2\alpha} - 1}{1 - 2\alpha} \left( \frac{T}{2} \right)^{1-2\alpha}. \quad (64)$$

Therefore, the right hand side of (60) is bounded above by

$$G\rho + C \sum_{t=T/2+1}^T t^{-2\alpha} \leq \left( G\rho + C \frac{1 - (1/2)^{1-2\alpha}}{1 - 2\alpha} \right) \cdot T^{1-2\alpha}. \quad (65)$$

Dividing the above term by the lower bound (62) to  $\sum_{t=T/2+1}^T \gamma_t$  yields (16).

## B.2 Convergence to stationary point of (1)

Recall that the set of stationary points to (1) is defined as:

$$\mathcal{C}^* := \{ \bar{\theta} \in \mathcal{C} : \max_{\theta \in \mathcal{C}} \langle \nabla F(\bar{\theta}), \bar{\theta} - \theta \rangle = 0 \}. \quad (66)$$

We state the following Nurminkii's sufficient condition:

**Theorem 3** [48, Theorem 1] *Consider a sequence  $\{\bar{\theta}_t\}_{t \geq 1}$  in a compact set  $\mathcal{C}$ . Suppose that the following hold<sup>4</sup>:*

- A1.  $\lim_{t \rightarrow \infty} \|\bar{\theta}_{t+1} - \bar{\theta}_t\| = 0$ .
- A2. *Let  $\underline{\theta}$  be a limit point of  $\{\bar{\theta}_t\}_{t \geq 1}$  and  $\{\theta_{s_t}\}_{t \geq 1}$  be a subsequence that converges to  $\underline{\theta}$ . If  $\underline{\theta} \notin \mathcal{C}^*$ , then for any  $t$  and some sufficiently small  $\epsilon > 0$ , there exists a finite  $s$  such that  $\|\bar{\theta}_s - \theta_{s_t}\| > \epsilon$  and  $s > s_t$ .*

---

<sup>4</sup>To give a clearer presentation, we have rephrased conditions A2 and A3 from the original Nurminkii's conditions.

A3. Let  $\underline{\theta}$  be a limit point of  $\{\bar{\theta}_t\}_{t \geq 1}$  and  $\{\theta_{s_t}\}_{t \geq 1}$  be a subsequence that converges to  $\underline{\theta}$ . If  $\underline{\theta} \notin \mathcal{C}^*$ , then for any  $t$  and some sufficiently small  $\epsilon > 0$ , we can define

$$\tau_t := \min_{s > s_t} s \quad \text{s.t.} \quad \|\bar{\theta}_s - \bar{\theta}_{s_t}\| > \epsilon \quad (67)$$

where  $\tau_t$  is finite. Also, there exists a continuous function  $W(\bar{\theta})$  that takes a finite number of values in  $\mathcal{C}^*$  with

$$\limsup_{t \rightarrow \infty} W(\bar{\theta}_{\tau_t}) < \lim_{t \rightarrow \infty} W(\bar{\theta}_{s_t}) . \quad (68)$$

Then the sequence  $\{W(\bar{\theta}_t)\}_{t \geq 1}$  converges and the limit points of the sequence  $\{\bar{\theta}_t\}_{t \geq 1}$  belongs to the set  $\mathcal{C}^*$ .

Our plan is to apply the above theorem to prove that every limit point of  $\{\bar{\theta}_t\}_{t \geq 1}$  are in  $\mathcal{C}^*$ . First of all, A1 can be easily verified since

$$\|\bar{\theta}_{t+1} - \bar{\theta}_t\| \leq \frac{\gamma_t}{N} \sum_{i=1}^N \|\mathbf{a}_t^i - \bar{\theta}_t\| \leq \frac{\gamma_t \bar{\rho}}{N} \quad (69)$$

and  $\gamma_t \rightarrow 0$  as  $t \rightarrow \infty$ .

As  $\mathcal{C}$  is compact, there exists a convergent subsequence  $\{\bar{\theta}_{s_t}\}_{t \geq 1}$  of the sequence of iterates generated by the DeFW algorithm. Let  $\underline{\theta}$  be the limit point of  $\{\bar{\theta}_{s_t}\}_{t \geq 1}$  and  $\underline{\theta} \notin \mathcal{C}^*$ . We shall verify A2 by contradiction. In particular, fix a sufficiently small  $\epsilon > 0$  and assume that the following holds:

$$\|\bar{\theta}_s - \bar{\theta}_{s_t}\| \leq \epsilon, \quad \forall s > s_t, \quad \forall t \geq 1 . \quad (70)$$

As  $\{\bar{\theta}_{s_t}\}_{t \geq 1}$  converges to  $\underline{\theta}$ , the assumption (70) implies that for some sufficiently large  $t$  and any  $s > s_t$ , we have  $\bar{\theta}_s \in \mathcal{B}_{2\epsilon}(\underline{\theta})$ , i.e., the ball of radius  $2\epsilon$  centered at  $\underline{\theta}$ .

Since  $\underline{\theta} \notin \mathcal{C}^*$ , the following holds for some  $\delta > 0$ ,

$$\langle \nabla F(\bar{\theta}_s), \theta - \bar{\theta}_s \rangle \leq -\delta < 0, \quad \forall \theta \in \mathcal{C}, \quad \forall s > s_t . \quad (71)$$

In particular, we have  $\langle \nabla F(\bar{\theta}_s), \bar{\mathbf{a}}_s - \bar{\theta}_s \rangle \leq -\delta$ , where we recall that  $\bar{\mathbf{a}}_s = \arg \min_{\mathbf{a} \in \mathcal{C}} \langle \nabla F(\bar{\theta}_s), \mathbf{a} \rangle$ .

On the other hand, from (58) and using H1, H2, it holds true for all  $t \geq 1$  that:

$$\begin{aligned} F(\bar{\theta}_{t+1}) - F(\bar{\theta}_t) &\leq \gamma_t \cdot \langle \nabla F(\bar{\theta}_t), \bar{\mathbf{a}}_t - \bar{\theta}_t \rangle \\ &\quad + \gamma_t \cdot \mathcal{O}(t^{-\alpha}) + \gamma_t^2 L \bar{\rho}^2 / 2 . \end{aligned} \quad (72)$$

To arrive at a contradiction, we let  $s > s_t$  and sum up the two sides of (72) from  $t = s_t$  to  $t = s$  and consider the following chain of inequality:

$$F(\bar{\theta}_s) - F(\bar{\theta}_{s_t}) \leq \sum_{\ell=s_t}^s \gamma_\ell (\langle \nabla F(\bar{\theta}_\ell), \bar{\mathbf{a}}_\ell - \bar{\theta}_\ell \rangle + \mathcal{O}(\ell^{-\alpha})) \leq -\delta \sum_{\ell=s_t}^s \gamma_\ell + \sum_{\ell=s_t}^s \gamma_\ell \mathcal{O}(\ell^{-\alpha}) , \quad (73)$$

where the first inequality is due to the fact that  $\gamma_\ell^2 L \bar{\rho}^2 / 2 = \gamma_\ell \mathcal{O}(\ell^{-\alpha})$  and the second inequality is due to (71). Rearranging terms in (73), we have

$$F(\bar{\theta}_s) - F(\bar{\theta}_{s_t}) - \sum_{\ell=s_t}^s C \cdot \ell^{-2\alpha} \leq -\delta \sum_{\ell=s_t}^s \ell^{-\alpha} , \quad (74)$$

for some  $C < \infty$ . As  $1 \geq \alpha > 0.5$ , we have  $\lim_{s \rightarrow \infty} \sum_{\ell=s_t}^s \ell^{-2\alpha} < \infty$  on the left hand side and  $\lim_{s \rightarrow \infty} \sum_{\ell=s_t}^s \ell^{-\alpha} \rightarrow +\infty$  on the right hand side. Letting  $s \rightarrow \infty$  on the both side of (74) implies

$$\lim_{s \rightarrow \infty} F(\bar{\theta}_s) - F(\bar{\theta}_{s_t}) < -\infty, \quad (75)$$

This leads to a contradiction to (71) since  $F(\theta)$  is bounded over  $\mathcal{C}$ . We conclude that A2 holds for the DeFW algorithm.

The remaining task is to verify A3. We notice that the indices  $\tau_t$  in (67) are well defined since A2 holds. Take  $W(\theta) = F(\theta)$  and notice that the image  $F(\mathcal{C}^*)$  is a finite set (cf. H3). By the definition of  $\tau_t$ , we have  $\bar{\theta}_s \in \mathcal{B}_\epsilon(\bar{\theta}_{s_t})$  for all  $s_t \leq s \leq \tau_t - 1$ . Again for some sufficiently large  $t$ , we have  $\bar{\theta}_s \in \mathcal{B}_\epsilon(\bar{\theta}_{s_t}) \subseteq \mathcal{B}_{2\epsilon}(\bar{\theta})$  and the inequality (73) holds for  $s = \tau_t - 1$ . This gives:

$$F(\bar{\theta}_{\tau_t}) - F(\bar{\theta}_{s_t}) \leq \sum_{\ell=s_t}^{\tau_t-1} \gamma_\ell \cdot (-\delta + \mathcal{O}(\ell^{-\alpha})). \quad (76)$$

On the other hand, we have  $\bar{\theta}_{\tau_t} \notin \mathcal{B}_\epsilon(\bar{\theta}_{s_t})$  and thus

$$\epsilon < \|\bar{\theta}_{\tau_t} - \bar{\theta}_{s_t}\| \leq \sum_{\ell=s_t}^{\tau_t-1} \gamma_\ell \left\| \sum_{i=1}^N \frac{\mathbf{a}_\ell^i}{N} - \bar{\theta}_\ell \right\| \leq \bar{\rho} \sum_{\ell=s_t}^{\tau_t-1} \gamma_\ell. \quad (77)$$

The above implies that  $\sum_{\ell=s_t}^{\tau_t-1} \gamma_\ell > \epsilon/\bar{\rho} > 0$ . Considering (76) again, observe that  $\mathcal{O}(\ell^{-\alpha})$  decays to zero, for some sufficiently large  $t$ , we have  $-\delta + \mathcal{O}(\ell^{-\alpha}) \leq -\delta' < 0$  if  $\ell \geq s_t$ . Therefore, (76) leads to

$$F(\bar{\theta}_{\tau_t}) - F(\bar{\theta}_{s_t}) \leq -\delta' \sum_{\ell=s_t}^{\tau_t-1} \gamma_\ell < -\frac{\delta' \epsilon}{\bar{\rho}} < 0. \quad (78)$$

Taking the limit  $t \rightarrow \infty$  on both sides leads to (68). The proof for the convergence to stationary point in Theorem 2 is completed by applying Theorem 3.

## C Proof of Lemma 1

For simplicity, we shall drop the dependence of  $\alpha$  in the constant  $t_0(\alpha)$ . It suffices to show that for all  $t \geq 1$ ,

$$\sqrt{\sum_{i=1}^N \|\bar{\theta}_t^i - \bar{\theta}_t\|_2^2} \leq \frac{C_p}{t^\alpha}, \quad C_p = (t_0)^\alpha \cdot \sqrt{N} \bar{\rho}. \quad (79)$$

We observe that for  $t = 1$  to  $t = t_0$ , the above inequality is true since  $\bar{\theta}_t^i, \bar{\theta}_t \in \mathcal{C}$  and the diameter of  $\mathcal{C}$  is bounded by  $\bar{\rho}$ . For the induction step, let us assume that  $\sqrt{\sum_{i=1}^N \|\bar{\theta}_t^i - \bar{\theta}_t\|^2} \leq C_p/t^\alpha$  for some  $t \geq t_0$ . Observe that

$$\theta_{t+1}^i = (1 - t^{-\alpha}) \bar{\theta}_t^i + t^{-\alpha} \mathbf{a}_t^i. \quad (80)$$

Using Fact 1, we observe that,

$$\sum_{i=1}^N \|\bar{\theta}_{t+1}^i - \bar{\theta}_{t+1}\|_2^2 \leq |\lambda_2(\mathbf{W})|^2 \cdot \sum_{j=1}^N \|(1 - t^{-\alpha})(\bar{\theta}_t^j - \bar{\theta}_t) + t^{-\alpha}(\mathbf{a}_t^j - \bar{\mathbf{a}}_t)\|_2^2, \quad (81)$$



where  $\tilde{\mathbf{a}}_t = N^{-1} \sum_{i=1}^N \mathbf{a}_t^i$  and we have used  $\bar{\boldsymbol{\theta}}_{t+1} = (1 - t^{-\alpha})\bar{\boldsymbol{\theta}}_t + t^{-\alpha}\tilde{\mathbf{a}}_t$ . The right hand side can be bounded by

$$\begin{aligned}
& \sum_{j=1}^N \|(1 - t^{-\alpha})(\bar{\boldsymbol{\theta}}_t^j - \bar{\boldsymbol{\theta}}_t) + t^{-\alpha}(\mathbf{a}_t^j - \tilde{\mathbf{a}}_t)\|_2^2 \\
& \leq \sum_{j=1}^N (\|\bar{\boldsymbol{\theta}}_t^j - \bar{\boldsymbol{\theta}}_t\|_2^2 + t^{-2\alpha}\bar{\rho}^2 + 2\bar{\rho}t^{-\alpha}\|\bar{\boldsymbol{\theta}}_t^j - \bar{\boldsymbol{\theta}}_t\|_2) \\
& \leq t^{-2\alpha}(C_p^2 + N\bar{\rho}^2) + 2\bar{\rho}t^{-\alpha}\sqrt{N} \sqrt{\sum_{j=1}^N \|\bar{\boldsymbol{\theta}}_t^j - \bar{\boldsymbol{\theta}}_t\|_2^2} \\
& \leq t^{-2\alpha}(C_p + \sqrt{N}\bar{\rho})^2 \leq \left( \frac{(t_0)^\alpha + 1}{(t_0)^\alpha \cdot t^\alpha} \cdot C_p \right)^2,
\end{aligned} \tag{82}$$

where we have used the boundedness of  $\mathcal{C}$  in the first inequality, the norm equivalence  $\sum_{j=1}^N |c_j| \leq \sqrt{N} \sqrt{\sum_{j=1}^N c_j^2}$  in the second inequality and the induction hypothesis in the third and fourth inequalities. Consequently, from (22), we observe that for all  $t \geq t_0$ ,

$$|\lambda_2(\mathbf{W})| \cdot \frac{(t_0)^\alpha + 1}{(t_0)^\alpha \cdot t^\alpha} \leq \frac{1}{(t+1)^\alpha}, \tag{83}$$

and the induction step is completed. Finally, Lemma 1 is proven by observing that (79) implies (24).

## D Proof of Lemma 2

We prove the first condition (27) using a simple induction. The condition is obviously true for the base step  $t = 1$ . For the induction step, suppose that (27) is true up to some  $t$ , we observe that

$$\sum_{i=1}^N \nabla_{t+1}^i F = \sum_{i=1}^N (\overline{\nabla_t^i F} - \nabla f_i(\bar{\boldsymbol{\theta}}_t^i)) + \sum_{i=1}^N \nabla f_i(\bar{\boldsymbol{\theta}}_{t+1}^i). \tag{84}$$

Note that the first term on the right hand side is zero due to the induction hypothesis. Thus, the induction step is completed and we conclude  $N^{-1} \sum_{i=1}^N \nabla_t^i F = N^{-1} \sum_{i=1}^N \nabla f_i(\bar{\boldsymbol{\theta}}_t^i)$  for all  $t \geq 1$ .

Then, we prove the second condition (28). For simplicity, we drop the dependence of  $\alpha$  in the constant  $t_0(\alpha)$ . Recall  $\bar{\nabla}_t F := N^{-1} \sum_{i=1}^N \nabla_t^i F$ . It suffices to prove:

$$\sqrt{\sum_{i=1}^N \|\bar{\nabla}_t^i F - \bar{\nabla}_t F\|_2^2} \leq \frac{C_g}{t^\alpha}, \quad C_g = 2\sqrt{N}(t_0)^\alpha(2C_p + \bar{\rho})L \tag{85}$$

for all  $t \geq 1$  using induction. For  $t = 1$  to  $t = t_0$ , the inequality can be easily proven using the boundedness of the gradients. For the induction step, we suppose that  $\sqrt{\sum_{i=1}^N \|\bar{\nabla}_t^i F - \bar{\nabla}_t F\|_2^2} \leq$

$C_g/t^\alpha$  for some  $t \geq t_0$ . Define the slack variable  $\delta f_{t+1}^i := \nabla f_i(\bar{\theta}_{t+1}^i) - \nabla f_i(\bar{\theta}_t^i)$  and observe that  $\overline{\nabla_{t+1}^i F} = \delta f_{t+1}^i + \overline{\nabla_t^i F}$ . Using Fact 1, we have

$$\sum_{i=1}^N \|\overline{\nabla_{t+1}^i F} - \overline{\nabla_{t+1} F}\|_2^2 \leq |\lambda_2(\mathbf{W})|^2 \cdot \sum_{i=1}^N \|\overline{\nabla_t^i F} + \delta f_{t+1}^i - \overline{\nabla_{t+1} F}\|_2^2, \quad (86)$$

Similarly, let us define  $\delta F_{t+1} := \overline{\nabla_{t+1} F} - \overline{\nabla_t F} = N^{-1} \sum_{i=1}^N \delta f_{t+1}^i$ . We can upper bound the right hand side of (86) as

$$\begin{aligned} \sum_{i=1}^N \|\overline{\nabla_t^i F} + \delta f_{t+1}^i - \overline{\nabla_{t+1} F}\|_2^2 &\leq \sum_{i=1}^N \left( \|\overline{\nabla_t^i F} - \overline{\nabla_t F}\|_2^2 + \|\delta f_{t+1}^i - \delta F_{t+1}\|_2^2 \right. \\ &\quad \left. + 2 \cdot \|\delta f_{t+1}^i - \delta F_{t+1}\|_2 \cdot \|\overline{\nabla_t^i F} - \overline{\nabla_t F}\|_2 \right), \end{aligned} \quad (87)$$

where the first inequality is obtained by expanding the squared  $\ell_2$  norm and applying Cauchy-Schwartz inequality.

Observe that for all  $i \in [N]$ , we have the following chain:

$$\begin{aligned} \|\delta f_{t+1}^i\|_2 &= \|\nabla f_i(\bar{\theta}_{t+1}^i) - \nabla f_i(\bar{\theta}_t^i)\|_2 \leq L \|\bar{\theta}_{t+1}^i - \bar{\theta}_t^i\|_2 \\ &\leq L \left\| \sum_{j=1}^N W_{ij} ((\theta_{t+1}^j - \bar{\theta}_t^j) + (\bar{\theta}_t^j - \bar{\theta}_t^i)) \right\|_2 \\ &\leq L \sum_{j=1}^N W_{ij} \left( t^{-\alpha} \bar{\rho} + 2C_p t^{-\alpha} \right) = (2C_p + \bar{\rho}) L t^{-\alpha}, \end{aligned} \quad (88)$$

where the last inequality is due to the convexity of  $\ell_2$  norm, the update rule in line 5 of Algorithm 1 and the results from Lemma 1. Using the triangular inequality, we observe that

$$\begin{aligned} \|\delta f_{t+1}^i - \delta F_{t+1}\|_2 &= \left\| \left(1 - \frac{1}{N}\right) \delta_{t+1}^i + \frac{1}{N} \sum_{j \neq i} \delta_{t+1}^j \right\|_2 \\ &\leq \left(1 - \frac{1}{N}\right) \|\delta_{t+1}^i\|_2 + \frac{1}{N} \sum_{j \neq i} \|\delta_{t+1}^j\|_2 \\ &\leq 2 \left(1 - \frac{1}{N}\right) (2C_p + \bar{\rho}) L t^{-\alpha} \leq 2(2C_p + \bar{\rho}) L t^{-\alpha}. \end{aligned} \quad (89)$$

Finally, applying the induction hypothesis, the right hand side of Eq. (87) can be bounded by

$$\begin{aligned} \sum_{i=1}^N \|\overline{\nabla_t^i F} + \delta f_{t+1}^i - \overline{\nabla_{t+1} F}\|_2^2 &\leq t^{-2\alpha} (C_g^2 + 4N(2C_p + \bar{\rho})^2 L^2) \\ &\quad + t^{-\alpha} 4L(2C_p + \bar{\rho}) \sqrt{N} \sqrt{\sum_{i=1}^N \|\overline{\nabla_t^i F} - \overline{\nabla_t F}\|_2^2} \\ &\leq t^{-2\alpha} \cdot (C_g + 2L\sqrt{N}(2C_p + \bar{\rho}))^2 \leq \left( \frac{(t_0)^\alpha + 1}{(t_0)^\alpha \cdot t^\alpha} \cdot C_g \right)^2, \end{aligned} \quad (90)$$

where we have used the fact that  $\sum_{i=1}^N \|\overline{\nabla_t^i F} - \nabla_t F\|_2 \leq \sqrt{N} \sqrt{\sum_{i=1}^N \|\overline{\nabla_t^i F} - \nabla_t F\|_2^2}$  in the first inequality above. Invoking (22), we can upper bound the right hand side of (86) by  $C_g^2/(t+1)^{2\alpha}$  for all  $t \geq t_0$ . Taking square root on both sides of the inequality completes the induction step. Consequently, (28) can be implied by (86).

## E Proof of Lemma 3

We begin the proof by applying the triangular inequality:

$$\begin{aligned} \left\| \xi_t^{-1} \overline{\nabla_t^i F} - \frac{1}{N} \sum_{j=1}^N \nabla f_j(\bar{\theta}_t^j) \right\|_\infty &\leq \xi_t^{-1} \cdot \left\| \overline{\nabla_t^i F} - \frac{1}{N} \sum_{j=1}^N \nabla f_j(\bar{\theta}_t^j) \odot \mathbf{1}_{\Omega_t} \right\|_\infty \\ &\quad + \left\| \left( \frac{1}{N} \sum_{j=1}^N \nabla f_j(\bar{\theta}_t^j) \right) \odot (\xi_t^{-1} \mathbf{1}_{\Omega_t} - \mathbf{1}) \right\|_\infty, \end{aligned} \quad (91)$$

where  $\mathbf{1}$  denotes the all-one vector.

For the first term in the right hand side of (91), observe that  $\overline{\nabla_t^i F}$  is obtained by applying the GAC updates on the sparsified local gradients  $\nabla f_i(\bar{\theta}_t^i) \odot \mathbf{1}_{\Omega_t}$  for  $\ell_t = \lceil C_l + \log(t)/\log |\lambda_2^{-1}(\mathbf{W})| \rceil$  rounds, applying Fact 1 yields the following for all  $i \in [N]$ :

$$\begin{aligned} &\left\| \overline{\nabla_t^i F} - \frac{1}{N} \sum_{j=1}^N \nabla f_j(\bar{\theta}_t^j) \odot \mathbf{1}_{\Omega_t} \right\|_\infty \\ &\leq |\lambda_2(\mathbf{W})|^{\ell_t} \cdot \left\| (\nabla f_i(\bar{\theta}_t^i) - \frac{1}{N} \sum_{j=1}^N \nabla f_j(\bar{\theta}_t^j)) \odot \mathbf{1}_{\Omega_t} \right\|_\infty \\ &\leq |\lambda_2(\mathbf{W})|^{C_l} \cdot B/t, \end{aligned} \quad (92)$$

for some  $B < \infty$  since the gradients are bounded.

For the second term in the right hand side of (91), we first apply the inequality  $\|(N^{-1} \sum_{i=1}^N \nabla f_i(\bar{\theta}_t^i)) \odot (\xi_t^{-1} \mathbf{1}_{\Omega_t} - \mathbf{1})\|_\infty \leq \|N^{-1} \sum_{i=1}^N \nabla f_i(\bar{\theta}_t^i)\|_\infty \|(\xi_t^{-1} \mathbf{1}_{\Omega_t} - \mathbf{1})\|_\infty$  from [49]. Now, the probability that co-ordinate  $k$  is included is given by:

$$P(k \in \Omega_t) = 1 - P\left(\bigcap_{i=1}^N k \notin \Omega_{t,i}\right) = 1 - \left(1 - \frac{1}{d}\right)^{p_t N} = \xi_t, \quad (93)$$

and that  $\mathbb{E}[\mathbf{1}_{\Omega_t}] = \xi_t \mathbf{1}$ . Then, observing that each element in  $\xi_t^{-1} \mathbf{1}_{\Omega_t}$  is bounded in  $[0, \xi_t^{-1}]$  and applying the Hoeffding's inequality [50], the following holds true for all  $x > 0$ :

$$P(\|\xi_t^{-1} \mathbf{1}_{\Omega_t} - \mathbf{1}\|_\infty \geq x) \leq 2d \cdot e^{-2x^2/\xi_t^{-2}}, \quad (94)$$

where we have applied a union bound argument to take care of the  $\ell_\infty$ -norm.

Setting  $x = \xi_t^{-1} \sqrt{(\log(2dt^2) - \log \epsilon)/2}$  and applying another union bound show that with probability at least  $1 - (\pi^2 \epsilon/6)$ , the following holds for all  $t \geq 1$ :

$$\begin{aligned} & \left\| \frac{1}{N} \sum_{i=1}^N \nabla f_i(\bar{\theta}_t^i) \odot (\xi_t^{-1} \mathbf{1}_{\Omega_t} - \mathbf{1}) \right\|_{\infty} \\ & \leq \xi_t^{-1} \left\| \frac{1}{N} \sum_{i=1}^N \nabla f_i(\bar{\theta}_t^i) \right\|_{\infty} \sqrt{\frac{\log(2dt^2/\epsilon)}{2}}, \end{aligned} \tag{95}$$

As  $d \gg 0$ , we have  $\xi_t^{-1} \approx d/(p_t N)$ . Recalling  $p_t = \Omega(t)$  yields the desired result in Lemma 3.

## References

- [1] J. Lafond, H.-T. Wai, and E. Moulines, “D-FW: Communication Efficient Distributed Algorithms for High-dimensional Sparse Optimization,” in *Proc ICASSP*, 2016.
- [2] H.-T. Wai, A. Scaglione, J. Lafond, and E. Moulines, “A projection-free decentralized algorithm for non-convex optimization,” in *Proc GlobalSIP*, December 2016.
- [3] V. Cevher, S. Becker, and M. Schmidt, “Convex Optimization for Big Data: Scalable, randomized, and parallel algorithms for big data analytics,” *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 32–43, Sep. 2014.
- [4] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, “Diffusion strategies for adaptation and learning over networks: an examination of distributed strategies and network behavior,” *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.
- [5] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
- [6] C. Ravazzi, S. M. Fosson, and E. Magli, “Randomized algorithms for distributed nonlinear optimization under sparsity constraints,” *IEEE Trans. on Signal Process.*, vol. 64, no. 6, pp. 1420–1434, Mar 2016.
- [7] S. Patterson, Y. C. Eldar, and I. Keidar, “Distributed compressed sensing for static and time-varying networks,” *IEEE Trans. on Signal Process.*, vol. 62, no. 19, pp. 4931–4946, July 2014.
- [8] J. Tsitsiklis, “Problems in decentralized decision making and computation,” Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, M.I.T., Boston, MA, 1984.
- [9] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, “Gossip Algorithms for Distributed Signal Processing,” *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.
- [10] S. S. Ram, A. Nedic, and V. V. Veeravalli, “A new class of distributed optimization algorithms : application to regression of distributed data,” *Optimization Methods and Software*, no. 1, pp. 37–41, Feb. 2012.
- [11] W. Shi, Q. Ling, G. Wu, and W. Yin, “A Proximal Gradient Algorithm for Decentralized Composite Optimization,” *IEEE Trans. on Signal Process.*, pp. 1–11, 2015.

- [12] T.-H. Chang, A. Nedic, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1524–1538, June 2014.
- [13] M. Hong, "Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications," *CoRR*, vol. abs/1604.00543, Apr 2016.
- [14] Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento, "A parallel stochastic approximation method for nonconvex multi-agent optimization problems," *CoRR*, vol. abs/1410.5076, Oct 2014.
- [15] P. D. Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Trans. on Signal and Info. Process. over Networks*, 2016.
- [16] X. Li and A. Scaglione, "Convergence and applications of a gossip based gauss newton algorithm," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5231–5246, Nov 2013.
- [17] E. Wei and A. Ozdaglar, "On the  $o(1/k)$  convergence of asynchronous distributed alternating direction method of multipliers," *CoRR*, 2013.
- [18] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Fast distributed gradient methods," *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1131–1146, May 2014.
- [19] J. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. Autom. Control*, vol. 57, no. 3, pp. 592–606, March 2012.
- [20] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, Feb 2013.
- [21] H.-T. Wai and A. Scaglione, "Consensus on state and time: Decentralized regression with asynchronous sampling," *IEEE Trans. on Signal Process.*, vol. 63, no. 11, pp. 2972–2985, June 2015.
- [22] H.-T. Wai, T.-H. Chang, and A. Scaglione, "A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning," in *Proc ICASSP*, Apr 2015, pp. 3546–3550.
- [23] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Res. Logis. Quart.*, 1956.
- [24] M. Jaggi and M. Sulovsky, "A simple algorithm for nuclear norm regularized problems," in *ICML*, 2010.
- [25] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with frank-wolfe algorithm," in *ECCV*, 2014.
- [26] L. Zhang, V. Kekatos, and G. B. Giannakis, "Scalable electric vehicle charging protocols," *CoRR*, 2016.

- [27] M. Fukushima, “A modified frank-wolfe algorithm for solving the traffic assignment problem,” *Transportation Research Part B: Methodological*, vol. 18, no. 2, pp. 169–177, April 1984.
- [28] M. Jaggi, “Revisiting Frank-Wolfe: Projection-free sparse convex optimization,” in *ICML*, 2013.
- [29] S. Ghadimi and G. Lan, “Accelerated gradient methods for nonconvex nonlinear and stochastic programming,” *Mathematical Programming*, vol. 156, no. 1, pp. 59–99, Feb 2015.
- [30] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, “Efficient projections onto the  $l_1$ -ball for learning in high dimensions,” in *ICML*. ACM, 2008, pp. 272–279, <http://www.cs.berkeley.edu/~jduchi/projects/DuchiShSiCh08.html>.
- [31] G. H. Golub and C. F. van Loan, *Matrix computations*, 4th ed. Johns Hopkins University Press, Baltimore, MD, 2013.
- [32] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, “Randomized gossip algorithms,” *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.
- [33] A. Defazio, F. Bach, and S. Lacoste-Julien, “Saga: A fast incremental gradient method with support for non-strongly convex composite objectives,” in *NIPS*, 2014.
- [34] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, “On distributed averaging algorithms and quantization effects,” *IEEE Trans. Autom. Control*, vol. 54, no. 11, pp. 2506–2517, Nov 2009.
- [35] Q. Ling, Y. Xu, W. Yin, and Z. Wen, “Decentralized low-rank matrix completion,” in *Proc ICASSP*, Mar 2012.
- [36] L. Mackey, A. Talwalkar, and M. I. Jordan, “Distributed matrix completion and robust factorization,” *Journal of Machine Learning Research*, vol. 16, pp. 913–960, 2015.
- [37] H.-F. Yu, C.-J. Hsieh, S. Si, and I. Dhillon, “Scalable coordinate descent approaches to parallel matrix factorization for recommender systems,” in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 765–774.
- [38] B. Recht and C. Ré, “Parallel stochastic gradient algorithms for large-scale matrix completion,” *Mathematical Programming Computation*, vol. 5, no. 2, pp. 201–226, 2013.
- [39] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Sparse and low-rank matrix decompositions,” in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*. IEEE, 2009, pp. 962–967.
- [40] G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten, “Fast Sparse Representation Based on Smoothed  $L_0$  Norm,” in *ICA*, ser. Lecture Notes in Computer Science. Springer, Sep. 2007, pp. 389–396.
- [41] H. Attiya and J. Welch, *Distributed Computing: Fundamentals, Simulations, and Advanced Topics*. Wiley, 2004.

- [42] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, Sep. 2004.
- [43] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *ACM TiiS*, Jan 2015.
- [44] E. v. Berg, M. P. Friedlander, G. Hennenfent, F. Herrmann, R. Saab, and Ö. Yilmaz, “Sparco: A testing framework for sparse reconstruction,” Dept. Computer Science, University of British Columbia, Vancouver, Tech. Rep. TR-2007-20, October 2007.
- [45] T. Blumensath and M. E. Davies, “Iterative hard thresholding for compressed sensing,” *CoRR*, May 2008.
- [46] J. Lafond, H.-T. Wai, and E. Moulines, “On the online frank-wolfe algorithms for convex and non-convex optimizations,” *ArXiv e-prints*, 2016. [Online]. Available: <http://arxiv.org/pdf/1510.01171v2.pdf>
- [47] B. P. Polyak, *Introduction to Optimization*. Optimization Software, Inc., 1987.
- [48] E. A. Nirminskii, “Convergence conditions for nonlinear programming algorithms,” *Cybernetics*, no. 6, pp. 79–81, Nov 1972.
- [49] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*. Cambridge: Cambridge University Press, 1994, corrected reprint of the 1991 original.
- [50] P. Massart, *Concentration Inequalities and Model Selection*. Springer, 2003.